ＤＤＢＪニュースレター　　No. 9　　1990年 5月

DNA Data Bank of Japan
国立遺伝学研究所
遺伝情報研究センター
遺伝情報分析研究室内

目　　次

List of terminal servers at computers from which DDBJ is accessible via Japanese Academic Internet.

DNA Data Bank of Japan
National Institute of Genetics

Organizations	Host name	bps	phone/others	(terminal server for DDBJ) loginname	password
Tokyo Univ. VAX 8600	tansei	64kbps	03-814-7271 7 bits even parity 2400,9600 bps	a88657	DDBJ-NIG
NIG Sun 4/390	ddbjs4s	64kbps	0559-75-6036,6037 DDX-P: (163-060)-522-5127 8 bit no parity 1200,2400 bps full hostname: ddbjs4s.ddbj.nig.ac.jp ip-address: 133.39.128.2 For remote-login at a site on the Internet % telnet ddbj.nig.ac.jp or % telnet 133.39.128.2	(direct login) ddbjnews	(no password)
Nagoya Univ. FACOM A600/UTS	nucc	9.6kbps	052-781-3442 7 bits even parity 2400 bps	a88657a	DDBJ-NIG
Kyuto Univ. FACOM M780/UTS	kuduts	9.6kbps	075-753-7478 8 bits no parity 2400 bps	a88657	DDBJ-NIG
Osaka Univ. work stations	ccsun01 ccews01 ccews08 ccews09	9.6kbps	06-876-3145 7bits even parity 1200,9600 bps	a88657a a88657a a88657a a88657a	DDBJ-NIG DDBJ-NIG DDBJ-NIG DDBJ-NIG
Kyushu Univ. FACOM M780/UTS	kyu-cc	9.6kbps	092-631-3278 8 bits no parity 2400 bps	a88657a	DDBJ-NIG

by Sanzo Miyazawa
sanzo.miyazawa@ddbj.nig.ac.jp

DDBJ 1989年度 活動報告

宮沢　三造、林田　秀宜

1. データ収集入力について

1-1. DDBJ Release 5 及び Release 6 のリリースについて

DDBJ Release 5 は 1989年 7月に Release 6 は1990年 1 月にリリースされた。

DDBJ Rlease	date	Entries	Bases	Lines	comments
6	01/90	496	841236	29682	
5	07/89	395	679378	23895	
4	01/89	302	535985	18713	
3	07/88	230	345850	13773	
2	01/88	142	199392	8238	
1	07/87	66	108970	4291	

1-2. Journal split and direct data submissions

従来データバンク（DDBJ/EMBL/GenBank）は互いにデータ収集の重複を避けるため収集対象の学術雑誌を分担してきた。例えば日本は日本で出版される雑誌を中心にその他 J. Gen. Virol. を担当してきた。データバンクにおけるデータ収集システムが雑誌の scan に基づいていたため、この雑誌の分担はよく機能してきた。しかし近年データのデータバンクへの提出を呼びかける雑誌が増えてきたため、著者からのデータ提出が 50 % を越えるようになると、このシステムの欠点が目だつようになった。データ提出を基礎にデータ収集の分担を計れば当然、地域分担が望ましいであろう。著者との連絡の便からも便利であろう。日本では多くの研究者が論文を海外雑誌に投稿するので、地域分担が特に望まれる。そこで DDBJ は 1989年 6 月に三島でデータバンク会議が開かれた折り、その点を強く指摘し理解を求めた。EMBL/GenBank も地域分担が将来の姿であることから、少しづつ地域分担に移行していくことに賛意を示した。しかし地域分担に移行するにはデータ入力処理が迅速でないと論文からも入力され重複する可能性が多く、入力されたデータを通常のリリースを待たずただちに担当データバンクに送付する必要があることが指摘された。そこで DDBJ は入力されたデータをただちに電子メールで送付することを約束し、地域分担の第一歩としてGenBank担当の雑誌に関し日本の研究者のデータは DDBJ に提出してもらうことにGenBankと合意に達し、1989年 10月よりデータ処理に着手した。 EMBL 担当のものはとりあえずこれまでと同じくデータを電子メールで転送することにし、EMBL 及びDDBJ の用意が整いしだい新システムに移行することになった。　現在 DDBJ へのデータ提出はDDBJ，GenBank，EMBL担当ものが各々約8エントリー/月である。データ分担の新システムへの移行により現時点で約8エントリー/月の増加となる。移行が進めばかなりの増加になるものと思われる。

Direct Submissions to DDBJ

Month/Year	DDBJ	Forwarded to GenBank	Forwarded to EMBL
01-06/88	1	0 (0)	2 (2)
07	0	0 (0)	2 (1)
08	1	1 (0)	0 (0)
09	1	1 (0)	2 (2)
10	0	5 (0)	12 (3)
11	1	1 (0)	1 (0)
12	3	9 (0)	8 (1)
01/89	6	10 (7)	6 (4)
02	2	2 (0)	11(10)
03	3	6 (1)	1 (0)
04	10	4 (2)	3 (0)
05	3	4 (0)	6 (0)
06	2	1 (0)	3 (1)
07	6	5 (1)	3 (0)
08	6	3 (0)	6 (0)
09	12	4 (4)	8 (5)

	DDBJ	Forwarded to EMBL
10	6	9 (1)
11	10	7 (0)
12	7	6 (0)
01/90		

* Numbers in parenthesis indicate the numbers of acknowledgements sent to ddbj@ddbj.nig.ac.jp

1-3. ネットワークデータベースサーバー；データベースは毎日更新

データバンクは多くの学術雑誌と著者へのDNA データ提出の呼掛けに関し協力関係にあるが、その際学術雑誌よりデータバンクは提出されたデータを一定の期間内に一般に利用可能にすることを条件として求められています。これは当然の要求であろうと思いますのでデータバンクもこの条件を満たせるよう努力しています。EMBL Data Library はそのためにネットワークファイルサーバーを稼働させ、研究者が電子メールにより最新のデータを得ることを可能にしています。 DDBJ も DDBJ が入力したデータを可能な限り素早く利用可能にするためネットワークデータベースサーバーを1989年 12月より稼働させています。日本でも序々に電子メール（Junet, Bitnet）が利用可能な大学、企業が増えてきましたのでネットワークサーバーの利用も増えるものと思います。

ネットワークサーバーで利用できるデータベースは GenBank, EMBL, DDBJ データベース及びSwissProt, PIR, PRF等の蛋白質データベースである。 GenBank, EMBL データベースはほぼ毎日入力されたエントリーが電子メールで送付されてくる。一日平均して、約400 KB（フロッピー0.5枚）程度のデータに及ぶ。データベースはこれら新データをただちに利用可能にするため 1 日に 2度更新されている。 DDBJ 入力データも同様である。しかしこれら最新のデータは通常のリリースに比べエラーも多いので承知願いたい。

このネットワークサーバーは、配列データの検索機能もあり、EMBL の提供するネットワークファイルサーバーより便利な機能を持っている。

Statistics of the use of NetServ@ddbj.nig.ac.jp

	jp	Internet	Bitnet	au	ca
12/89	3	2	14	1	1
1/90 (1/19)	1	3	1	0	4

1-4. 関係データベース（RDB）への移行スケジュール

データバンクは現在のフラットファイルによるデータベースから関係データベースへの移行を共同で推進している。この関係データベースへの移行はデータ管理を容易にし又今後増加するであろうゲノムデータの管理を可能にします。

GenBank は関係データベースのためのスキーマのデザインを終え、トランザクション処理のためのソフトウエア及びデータ入力、更新のためのフロントエンドソフトウエアの開発ほぼ終え、1989 年 9月よりテスト期間、1990年 1月より RDB に完全に移行する計画である。EMBL は 1990年中に、DDBJ もGenBank の協力を得て1990年の前半には移行する計画である。 データベース管理ソフトウエアとしては、GenBank、DDBJ はSybaseを、EMBLはOracleである。GenBank、EMBL は異なるデータベース管理システムを採用しているが、一定の手順にしたがって、データベースへの追加、変更に関するデータを交換することにより実質的に同一のデータベースを構築する。

1-5. New feature table使用のスケジュール

現在磁気テープで配布されている DNA データベースの利用者は、関係データベースへの移行により配布データはどのようになるか知りたいであろう。配布されるデータのフォーマットは現在と同じものがサポートされる予定である。ただ、現在のデータの注釈部分（feature table）は、より計算機処理が可能なように、明確に定義されたキーワードを用い、また書式も複雑な DNA 遺伝情報を的確に表現できるようデザインされた新しいfeature tableに 1990年中に移行する計画である。新feature tableの定義マニュアルは 1988年 10月に完成したが、移行に伴うソフトウエア開発を減らすため関係データベースに移行後に採用することになった。

1-6. 研究者自身によるデータ入力のための支援ソフトウエア（AuthorIn）について

現在研究者のよるデータ提出では、配列データの注釈を自由書式により記述してもらっている。そのため一定の書式にのっとりデータベース化する際、専門知識が要求され入力のボトルネックになっている。このような状況を打破すべく研究者自身によるデータ入力を支援するソフトウエア、パーソナルコンピューターの上で稼働するソフトウエア AuthorIn を GenBank (IntelliGenetics)が開発している。テスト版（IBM-PC版）は 1989年 7月に完成し、実際の使用は 1990 年からを予定している。DDBJ はこのプログラムをNEC-PC9801 に移植する計画である。

このソフトウエアにおいては、フォームを完成する際ほとんどの項目においてHelp機能が利用でき、また多くの項目において候補のメニュウーが提示され適するものを選択することができる。このような方式を採用することにより用語を統一でき、また可能なかぎりのエラーチェックが可能となる。

1-6. CD-ROMについて

EMBL Data Library とGenBank(IntelliGenetics)は各々 CD-ROM 版のデータベースのリリースを1989年の夏頃から開始した。EMBL版は、通常のEMBLフォーマットと CD-ROM フォーマットで EMBLとSwissProtのデータベースを含む。GenBank 版は同じく通常のGenBankフォーマットと CD-ROM フォーマットで GenBank とPIRデータベースを含む。 EMBL 版には企業から提供されたIBM-PC用の検索プログラムが含まれている。 研究者からの要望が多いのでDDBJ はNEC-PC98で稼働するCD-ROM のための検索プログラムを開発することを計画している。 しかしCD-ROM をDDBJ が配布するかどうかは未定である。

2. データバンク会議報告

データバンク（DDBJ/EMBL/GenBank）は共同研究を推進するため毎年スタッフの会合を回り持ちで開催している。DDBJ は 1987年 11月の会合から参加している。1989 年度は DDBJ 担当ということで 1989年 6月に三島で開催された。

DDBJ はこの会議で、(1)データバンクに提出されたデータに関しては雑誌分担によらず入力処理することを提案した。また、(2) 現在EMBL またはGenBank データベースの利用者のほとんどは、これらデータベースはEMBL、GenBank により入力されたものと認識してい

ると思われるので、データベースに含まれる DDBJ 入力分に関しては少なくともリリースノートに明記することを求めた。前者の提案は認められ、DDBJ は GenBankの担当雑誌の投稿される日本の研究者のデータを1989年 10月から入力することになった。今後対象を EMBL にも拡大する予定である。後者の要望については、個々のエントリーに入力処理したデータバンクを明記するためのレコードを新たに追加することになった。また、この新レコードが追加されるまでは、accession number が D ではじまるエントリーは DDBJ 入力であることをリリースノートで明記することになった。

詳しい会議の報告は別紙を参照願いたい。

3. DNA Databases and Genome Projects 研究集会報告

データバンクの会合が三島で開かれたのを機に GenBank、EMBLデータベースの現状と将来計画をめぐって研究集会を開いた。またデータバンクとしては日本におけるゲノム解析計画の現状について知りたいとの考えから幾人かの研究者に現状を報告していただきました。多くの方に参加いただき有難うございました。研究集会のプログラムを参考のため付与します。

4. DDBJ 計算機環境

4-1. データベース検索・解析ソフトウエア；ネットワークサーバー機能と
電子メールによるデータベースの自動更新機能の追加

データベース検索・解析ソフトウエア FLAT にコマンドを電子メールで送付することによりデータベースを検索することが可能なネットワークデータベースサーバー機能と電子メールによるデータベースの自動更新機能が追加されました。ネットワークサーバー機能については付与した簡易マニュアルを参照ください。

FLAT は version 1.0 が1989年 5月に一般にリリースされ、UNIXシステムの上で利用を望む方に配布されていますが、この新版も1990年 3月頃にはリリースされる予定です。 この新版では、付随するFASTA プログラムも更新されています。

4-2. DDBJ/EMBL/GenBank new databases; 毎日更新

GenBankデータベースは1989年 9月から、また EMBL は 1990年 1月より毎日入力された最新のデータが電子メールで送られてきています。これらのデータは通常のリリースの後に入力された最新のものですが、通常のリリースに含まれるものと比べエラーチェックが十分でなく、また更新データも含むので、通常のリリースとは別に管理しています。これらのデータは new sequence databasesとして毎日 2 度更新されています。DDBJ も同様です。これらのデータベースの使用方法は % man flat でマニュアルを参照ください。

4-3. HGMLの利用の機会を提供

Yale Univ. 計算機を用い公開されている Human Genome Mapping Library を使用したいとの要望が高いので DDBJ 計算機からネットワーク経由で Yale Univ. の計算機（yalevm.ycc.yale.edu）を使用し HGML にアクセスできるようにしました。 HGML の使用料金は Howard Hughes Medical Instituteで負担します。

利用の方法を知るには計算機の上で getinfoコマンドを用いてください。

4-4. On-line version of Mendelian Inheritance in Man が利用可能になる予定。

Dr. Victor McKusickにより維持されている人遺伝病のデータベース でJohns Hopkins Univ. でオンラインサービスされている OMIM が、遺伝研藤山氏の御努力で近々 DDBJ 計算機で利用できるようになります。

4-5. C. elegansマップデータベースが利用可能になる予定。

遺伝研究所の小原氏の御努力で （MRC）により 構築されているデータベースがDDBJ 計算機で1990年 4月頃より利用可能になる予定です。

4-6. USENET 分散型電子ニュースシステムにおける 生物学者のためのフォーラム、BIONET、ニュースグループの購読及び JUNET 内への配布

米国にはUSENET という分散型電子ニュースシステムがあり計算機ネットワーク上で多くの科学者のフォーラムとして役立っている。USENETニュースはいくつかのニュースグループからなり分子生物学関係のニュースグループとしては BIONET ニュースグループが有名である。BIONET ニュースグループはNIH のサポートによるBIONET projectにより維持されていたが BIONET project の終結により現在では GenBank活動の一部としてサポートされている。BIONET ニュースは USENET だけでなく 電子郵便ネットワーク（例えばBITNET ネットワーク）にもメーリングリストの方式により転送されているので、購読者はほぼ全世界にわたっており、世界の研究者のフォーラムとして機能している。

そこで DDBJ は日本の研究者もこのニュースグループに参加できるよう BIONET ニュースの購読を1989年 6 月よりDDBJ 計算機システムで開始した。また日本中の他の計算機でも利用できるよう日本における類似のネットワーク JUNETへの転送も開始した。ニュースの投稿は自由ですので参加者が増えることを望んでいる。

またDDBJ 計算機システム以外には配布されないローカルのニュースグループとして DDBJ がある。DDBJ 利用者のフォーラムとして投稿を期待している。DDBJの活動も BIONETとDDBJ ニュースグループを利用しお知らせしている。

4-7. 国際ネットワーク（Internet）へ 1990年 3月末に接続

データバンク間で分散データベース構築のために、GenBank, EMBL, DDBJ間の計算機が高速ネットワークで接続されることが必要である。このため DDBJ は 2 年程前より関係機関(米国エネルギー省等）と連絡をもちESNET (Energy Science Network)経由で米国のInternet へ1989 年度に DDBJ 計算機を接続することを計画してきましたが、Hawai Univ. 経由で Internet が東大理学部まで延長されましたので、計画を変更しInternetに直接接続することにしました。 1990年 3月 28日に 64 kb で接続される予定です。

Internet には 1986年度で全米の約 2000のノードが 56 kb の高速通信回線で接続されています。このネットワークへの接続は米国の計算機で利用が公開されている関連データベースへのアクセスを DDBJ 計算機の上で可能にする意味でも有用であると思います。またこのネットワークへ加入する計算機が日本でも増えてきましたので遠隔地から DDBJ 計算機へのアクセスも容易になると思われます。また東大の大型計算機センターも接続されていますので、大学間ネットワークからのアクセスも可能になると思われます。（現在、東大大型計算機センター日立の計算機から直接 DDBJ 計算機を利用できるようになっていますが、N1 ネットワークからの利用は日立のソフトウエアの関係で、日立の計算機以外で

は接続ができません。）利用の詳細は計算機の上で getinfoコマンドを用いてください。

5．ニュースレター、ソフトウエア配布及びDNA、蛋白質データベースの配布に関して
資料参照下さい。

DNA Data Bank of Japan

DNA Database

Release 6, January 1990

496 entries, 841236 bases, 29682 lines

All entries included in this release have been created by DDBJ and will be incorporated into future releases of EMBL and GenBank databases. Entries created by DDBJ in the EMBL and GenBank databases may be identified, because their accession numbers begin with the letter "D".

Prepared by:

Sanzo Miyazawa	Manager/Database Administrator
Hidenori Hayashida	Scientific Reviewer
Yukiko Yamazaki/Eriko Hatada/Hiroaki Serizawa	Annotators/reviewers
Motono Horie/Shigeko Suzuki/Yumiko Sato	Secretaries/typists

DNA Data Bank of Japan
National Institute of Genetics
Center for Genetic Information research
Laboratory of Genetic Information Analyses

1111 Yata
Mishima, Shizuoka 411
Japan

Phone: +81 559 75 0771 x647
FAX: +81 559 75 6040
E-mail: ddbj@ddbj.nig.ac.jp

Files included:

1) ddbjrel.txt: release notes; this document
2) ddbjsub.txt: a data submission form
3) netserv.man: a brief manual of network DB server
4) ddbj.dna: DNA data
5) ddbj.cds: peptide coding sequences extracted from ddbj.dna
6) ddbj.pep: peptide sequences translated from ddbj.cds
7) ddbj.dir: a short directory file
8) ddbj.acc: an accession number index file
9) ddbj.aut: an author name index file
10) ddbj.jou: a journal name index file

The 5th and 6th files above were generated by using "seqext" and "pepttr" programs made by Dr. Jim Fickett in GenBank, Los Alamos National Laboratory.

The index files from the 7th through the 10th file above may include longer lines than 80 characters. If distribution tapes are written in the fixed size records of 80 bytes, longer lines than 80 characters will be folded in those files.

Acknowledgements:

We thank GenBank for helps, especially for providing us with such tools that are useful for quality control in data entry.

Appendix 1. The DNA sequence file, ddbj.dna.

The format of the DNA sequence file is the same as one of GenBank data which is described in the release note of GenBank Release 62.

Example:

```
------------------------------------------------------------------------------
LOCUS       BOVCYT2       269 bp ds-DNA             MAM       31-OCT-1989
DEFINITION  Bovine cytochrome P-450(11beta) gene, exon 2.
ACCESSION   D00450
KEYWORDS    cytochrome; cytochrome P-450; steroid hormone.
SEGMENT     2 of 8
SOURCE      Bovine adrenal DNA, clone lambdaB11beta(7-1).
  ORGANISM  Bos taurus
            Eukaryota; Metazoa; Chordata; Vertebrata; Tetrapoda; Mammalia;
            Eutheria; Artiodactyla.
REFERENCE   1  (bases 1 to 269)
  AUTHORS   Hashimoto,T., Morohashi,K. and Omura,T.
  TITLE     Cloning and characterization of bovine cytochrome P-450(11beta)
            genes
  JOURNAL   J. Biochem. 105, 676-679 (1989)
  STANDARD  full staff_entry
COMMENT
FEATURES        from  to/span     description
    pept    /    57  /   212     cytochrome P-450(11beta) gene, exon 2
    refnumbr     58       60     numbered codon 81 in [1]
BASE COUNT       49 a     74 c    110 g     36 t
ORIGIN
        1 agcggggacc aggcaggaca cggcccctca gcagaccgca gagctctgcc ccacaggtac
       61 gacgtgggag ggagacacat ggtgttcgtg atgctgcccg aggacgtgga gaggctgcag
      121 caggcggaca gccatcaccc tcagcggatg atcctggagc cctggctggc ctaccgacag
      181 gctcgcgggc acaagtgtgg cgtgttcttg ctgtgagggc cggtgggggt gcggagggag
      241 gggcagggcg gcctgggtgc agagagcac
//
------------------------------------------------------------------------------
```

Appendix 2. The peptide coding sequence file, ddbj.cds, extracted from ddbj.dna.

Example:

```
------------------------------------------------------------------------------
ID          BOVCYT2
            cytochrome P-450(11beta) gene, exon 2
LOCUS       BOVCYT2       269 bp ds-DNA             MAM       31-OCT-1989
ACCESSION   D00450
  ORGANISM  Bos taurus
            Eukaryota; Metazoa; Chordata; Vertebrata; Tetrapoda; Mammalia;
            Eutheria; Artiodactyla.
FEATURES        from  to/span     description
    pept    /    57  /   212   2 cytochrome P-450(11beta) gene, exon 2
COMPLETE    5':n  3':n
LENGTH      156
ORIGIN
        1 gtacgacgtg ggagggagac acatggtgtt cgtgatgctg cccgaggacg tggagaggct
       61 gcagcaggcg gacagccatc accctcagcg gatgatcctg gagccctggc tggcctaccg
      121 acaggctcgc gggcacaagt gtggcgtgtt cttgct
//
------------------------------------------------------------------------------
```

Appendix 3. The peptide sequence file, ddbj.pep, translated from ddbj.cds.

Example:

```
ID          BOVCYT2
            cytochrome P-450(11beta) gene, exon 2
LOCUS       BOVCYT2       269 bp ds-DNA              MAM       31-OCT-1989
ACCESSION   D00450
  ORGANISM  Bos taurus
            Eukaryota; Metazoa; Chordata; Vertebrata; Tetrapoda; Mammalia;
            Eutheria; Artiodactyla.
FEATURES        from  to/span     description
    pept        /    57  /   212    2 cytochrome P-450(11beta) gene, exon 2
COMPLETE    5':n  3':n
LENGTH      51
ORIGIN      Translated using phase 2
        1 ydvggrhmvf vmlpedverl qqadshhpqr milepwlayr qarghkcgvf l
//
```

Appendix 4. The short directory file, ddbj.dir.

Example:

```
Entry name #Acc             bp     Definition
------------------------------------------------------------------------
ACAADH1    D90004 ds-DNA    2467 + Acetobacter aceti(K6033) alcohol ...
ACH5SRR    D00106 ss-rRNA    114   Achromobacter xylosoxidans GIFU 543 ...
ACH5SRRX   D00107 ss-rRNA    120   Achromobacter cycloclastes IAM 1013 ...
AD4HEX     D00357 ds-DNA    3600   Adenovirus type41 Hexon gene, complete ...
AFA5SRR    D00105 ss-rRNA    118   Alcaligenes sp. NCIB 11015 ...
AFA5SRRA   D00104 ss-rRNA    115   Alcaligenes faecalis ATCC 8750 ...
AFA5SRRI   D00103 ss-rRNA    120   Alcaligenes faecalis IAM 1015 ...
ANITPI     D00013 ds-DNA    1900   Aspergillus nidulans triosephosphate ...
...
------------------------------------------------------------------------
                                   '+' means new entries.
```

Appendix 5. The accession number index file, ddbj.acc.

Example:

```
#Acc      Entry name #Primary accession
-------------------------------------------
D00001 -> ECOPBPA    D00001
D00002 -> ECOPYRC    D00002
D00003 -> HUMCYP450  D00003
D00004 -> FLBPB1L40  D00004
D00005 -> IBVMEM682  D00005
...
X16017 -> RDVAS5     D90033
-------------------------------------------
```

Appendix 6. The author name index file

Example:

```
Author    Entry name #Accession
---------------------------------------
Abe,M. -> GOTGH      D00476
Abouhaidar,M.G. -> CYMVCP     D00485
Abouhaidar,M.G. -> MTSCA      D00341
Abouhaidar,M.G. -> PMPCP      D00240
...
---------------------------------------
```

Appendix 7. The journal name index file

Example:

```
Jounral name                                    Entry name #Accession
-------------------------------------------------------------------
Agric. Biol. Chem. 47, 441-444 (1983) -> BOVCAS     D00412
Agric. Biol. Chem. 49, 2725-2731 (1985) -> MPOCPUP    D00379
Agric. Biol. Chem. 49, 2829-2831 (1985) -> HUMLYZ     D00413
Agric. Biol. Chem. 50, 2155-2158 (1986) -> ECONANA    D00067
...
-------------------------------------------------------------------
```

Report on Collaborative Meeting of DDBJ, The EMBL Data Library, and GenBank

at the NIG, Mishima, Japan

19-23 June 1989

Participants

EMBL Data Library		Graham Cameron	6/19 - 6/23
		Patricia Kahn	6/19 - 6/22
		David Hazledine	6/19 - 6/23
GenBank			
	IntelliGenetics	David Benton	6/19 - 6/23
		Julie Ryals	6/19 - 6/23
	LANL	Christian Burks	6/19 - 6/23
		Paul Gilna	6/19 - 6/23
		Michael Cinkosky	6/19 - 6/23
NIH		Jane Peterson	6/19 - 6/23
DDBJ		Sanzo Miyazawa	6/19 - 6/23
		Hidenori Hayashida	6/19 - 6/23

Agenda

1) Agenda	6/19	9:15- 9:30
2) Restructuring of databases Chair persons: Michael Cinkosky	6/19	9:30-16:00
2-a) Present status and schedule of RDB implementation Report from LANL by Michael Cinkosky. Report from EMBL by Dave Hazledine. Report from DDBJ by Sanzo Miyazawa.		
2-b) Data exchange Transaction protocol definition which is used by GenBank EMBL's standpoint		
3) Pattent data Chair person: Graham Cameron	6/19	17:00-18:00
4) Unification of nucleotide sequence databases Chair person: Graham Cameron	6/20	11:00-17:30
5) EMBNet (European Molecular Biology Network) report by Graham Cameron	6/21	9:15-10:45
6) Authorin and data submission form Chair person: Dave Benton	6/21	11:15-14:30
7) Tracking and documenting published and unpublished citations. Chair person: Christian Burks	6/21	14:30-17:30
8) Journal interactions Chair person: Patricia Kahn.	6/22	9:15-10:15
9) Problems with direct data submissions Chairperson: Patricia Kahn	6/22	10:15-15:00
10) Features table implementation schedule. Chair person: Paul Gilna.	6/22	16:00-17:30
11) New line type Chair person: Christian Burks	6/23	9:15-10:30
12) CD-ROM progress and joint use by EMBL, GenBank, PIR Chair person: David Benton	6/23	10:30-12:30
13) Conclusion Next meeting: place, date...	6/23	14:30-15:30

* coffee break: 10:45 - 11:00 and 15:30 - 16:00
* lunch: 12:30 - 14:00

1. Restructuring of databases at Genbank, EMBL, and DDBJ

1-1. GenBank by Michael Cinkosky

1) Status report and schedule of RDB implementation - genbank view

Michael Cinkosky reported on the progress GenBank has made on converting to the database to relational form. He demonstrated the Annotator's Workbench (the interactive interface that annotators will use to maintain the data after the conversion is complete) by logging in the the Los Alamos computers remotely from DDBJ. He explained that the GenBank staff expects to be able to produce release 63.0 (March 1990) from the relational database, including the new feature table format.

2) Data exchange - genbank's transaction protocol

Mr. Cinkosky reported that software for converting GenBank from flatfile form into the transactions (using the Transaction Protocol) was running, as was the code to execute these transactions into the relational database. He also indicated that the first version of the software for automatically processing transactions received by e-mail from authorin was running and was in regular use by the staff at Intelligenetics for testing the output of Authorin.

1-2. Progress on the RDBMS conversion at EMBL by David Hazledine

EMBL envisages the RDBMS conversion as a 2-stage process. Stage 1 consists of 4 tasks:

(a) complete the logical database design
(b) build software to load flat-files into the RDBMS
(c) build software to extract flat-files out of the RDBMS
(d) load the database into the RDBMS

When this is complete, all of our data will be stored in the RDBMS, but we will still be using the current data input/update software, which operates on flat-files rather than directly on data stored in the RDBMS. Stage 2 consists of the development of a new generation of data management tools which will operate directly on the RDBMS.

Tasks (a),(b),(c) of stage 1 are almost complete. The only data items which we cannot load and unload are now accession numbers and citations. We hope that stage 1 will be complete before November 1989, so that release 21 will be the first one which is generated by extracting flat-file entries out of the RDBMS.

We have not yet begun to design the software required for stage 2, and do not expect to complete the development of new tools before the second half of 1990 at the earliest.

1-3. Present status of DDBJ in the shift to RDB By sanzo Miyazawa

The shift from a flat DB to RDB is significantly delayed at DDBJ.
Last fiscal year, we submitted a grant proposal to purchase RDBMS, but we failed to get a grant. However, it is almost certain that our proposal will be granted this fiscal year. So, we decided to purchase RDBMS by diverting a different grant.

We decided to use SYBASE RDBMS, because of following reasons.

1) In general, using the same RDBMS as GenBank or EMBL uses makes it feasible or easy to build and maintain an identical database at multiple sites.

2) DDBJ does not have human resource enough to develop software for RDB by itself.

So, we must use either ORACLE or SYBASE RDBMS. Then, our choice of SYBASE was made because several databanks and organizations have decided to use SYBASE.

We will get SYBASE within a couple of months, probably by September. GenBank agreed that DDBJ became one of remote sites installing their software such as the data access library and interface library. So, we will try to install those software soon after we get SYBASE; I assume that GenBank is ready to release the software to us.

2. Pattent data

by Graham Cameron

The three major patent offices, that is the American Patent Office, the European Patent Office and the Japanese Patent Office note that an increasing amount of sequence data appears in patents. They have a real need to search these data in a database and they see the EMBL, Genbank and DDBJ collections as the appropriate place for the information. We in the databanks on the other hand are keen to see these data included in the interests of completeness for our users. Agreements about data exchange between the patent offices are nearly finalised and it is their suggestion that each of the patent offices collaborate with their corresponding database. Information on the development of these interactions, in three groups, was exchanged at the meeting.

GENBANK

Initial contacts have been made but further action from the Patent Office is awaited.

EMBL

Discussions of a possible contract between EMBL and the European Patent Offices are underway. EMBL would, for a fee, enter patent data and make all data available to the European Patent Office for search. Their copy of the database would probably be updated daily using the EMBNet system. EMBL has by way of experiment, entered and annotated some 20 patents. No major representational problems were encountered although the details of patent citation adds some new requirements. Patents which discuss families of sequences are in principle extremely difficult to represent but the European Patent Office staff were of the opinion that the pragmatic solution of storing only examples discussed in the application was not only satisfactory but probably desirable. Experiences with chemical databases in which they have attempted to fully represent the analogous concept of families have been very unsatisfactory.

DDBJ

As yet no contact has been made.

3. Unification of nucleotide sequence databases

By Graham Cameron

The need to unify the content of the various nucleotide sequence databases has long been recognised. Up until now we have attempted to ensure the unity of content by exchanging the data in the flat file format used for distribution. This has been problematic for a number of reasons hinging round (1) difficulties in automatically determining correspondences between information in different collections and (2) the propagation of updates.

Achieving identity of content of the various databases at any point, is only a worthwhile goal if mechanisms exist to maintain the databases identically. The GenBank transaction protocol aims to do this. It is a protocol designed to ensure that the same changes are applied to independent copies of the same information. At the meeting various approaches to ensuring identity of content

of the databases were discussed. The current model is one of concurrent copies of identical databases, that is to say no one copy is definitive. It was noted that, in terms of system design, this is an extremely unsatisfactory way of maintaining information, and that it would be beneficial to have one master copy of the database. This could be centralised at one site or perhaps distributed, with some parts of the information residing different sites. The present model utilising ad hoc methods for transmitting information about updates to the databases between the sites was seen as unsatisfactory. The transaction protocol attempts to formalise this in a way which ensures systematic propagation of updates.

There was an extensive discussion about the possibility of a genuinely unified single database and the desire to discuss this further in future collaborative meetings was expressed. Graham Cameron volunteered to pursue this further and to solicit and assemble the ideas of the various collaborators.

4. Authorin and data submission form By David Benton

David Benton reported that Authorin is in final qualification testing. Limited release (to 13 sites selected by LANL and about 85 investigators who have requested the program) will begin as soon as the program passes final qualification testing. The initial recipients will be asked to evaluate the program for both bugs and design limitations and flaws. A survey form designed for this purpose was distributed for comment. The Authorin Manual was also distributed for comment. The initial release phase will be limited to approximately 200 sites over the next two to three months. During this period we expect to be able to honor individual requests for the program in addition to the sites already identified. No bulk requests (including that from IRL Press's U.S. office) will be honored during the limited release phase. EMBL will coordinate with the IRL's European office to ensure that uniform policies are implemented for all NAR authors. Graham Cameron volunteered to identify non-native speakers of English who would be willing to participate in the evaluation of Authorin. Feedback from this release phase will be used as the basis for a maintenance release, currently scheduled for late September, and as additional requirements for release 2 of the program.

A list of program enhancements proposed by the beta testers of Authorin Release 1 was distributed and comments regarding priorities were requested.

Some discussion of future (following the limited release phase) mechanisms of Authorin distribution took place. The size of the Authorin Manual makes it unlikely that many journals will be willing to ship the entire distribution package. Several options were discussed:

1. Journals distribute business-reply request cards (with the hardcopy data submission forms). Using these cards, authors could request the program from either GenBank or the most convenient database site.

2. Journals distribute Authorin disks, installation instructions, quick reference card, and reply cards with which authors could request copies of the manual.

3. Journals simply instruct authors to contact the nearest database site to request either the program or a data submission form.

Christian Burks reported on the plans at LANL to develop a machine-parseable machine-readable data submission form, based on the current machine-readable data submission form. The feasibility of parsing such a form is under investigation at this time. If the decision to proceed is made, it is expected that changes to the machine-readable form will be required. Once these changes are made, it is proposed that the machine-readable form be maintained (by Patricia Kahn) in parallel with the hardcopy form and subject to the same revision procedures as are used for

the hardcopy form. Explicitly, this means that any modification to the machine-readable form must be approved by GenBank, EMBL, DDBJ, and PIR-International. At present the machine-readable data submission form is derived from the hardcopy form which is definitive. If a parseable machine-readable form is developed, its maintenance and revision would be semi-independent of the hardcopy data submission form. The next release of the hardcopy form is scheduled for November. It was not anticipated that the parseable machine-readable form would be available for distribution for comment prior to that release.

5. Tracking and documenting published and unpublished citations. by Christian Burks

All our data originally came in through the following path:

scientist -> journal articles -> library -> manual data entry -> database

This made the notion of what a citation was extremely unitary (data passed from one database to another was considered a transparent reflection of the original pathway outlined above).

However, DDBJ, EMBL, and GenBank are all facing increasing complexity in describing and tracking the source of their data. For example, the journal article is now most often <not> the direct source of the data coming into the databases, which gives exemplifies the need to tease apart the notion of "associated citation" and "data source".

This session was spent in identifying a number of the complexities that need to be better tracked:

(1) did the data come directly from the author, from in-house data entry from a journal articles, from a collaborating database, or from another third party?

(2) did the data get incorporated verbatim (automatically) from the the scientist's input, or were changes made or value added by intermediate sources and/or the receiving database?

(3) were the data published or not?

In addition, it became clear that tracking the dates associated with the various data path milestones would be highly desirable.

6. Data submissions and journal interactions By Patricia Kahn

The journal split for submissions (i.e. each database only processes those submissions relating to journals which it scans) was discussed extensively. It currently causes major problems because of the delay & extra work introduced into submission processing when a submission is received at the "wrong" database. Furthermore, DDBJ expressed the view that they would like to process all submissions originating in Japan. All three groups felt that the direct submission scheme should be opened up geographically so that authors may submit their data to any database, regardless of where they intend to publish, and the receiving database would assign an accession number and create an entry (or entries) from the submission.

One question this raised was whether such a change would cause a major shift in the balance of work among the databases. While this needs to be looked at more carefully, a quick-and-dirty analysis of submissions to EMBL and GenBank by country of origin suggested that there would not be major changes for these two groups, while DDBJ would experience a major rise in the number of submissions (but would like to go ahead anyway).

Several points arise from this: (1) since submissions represent database additions rather than updates, the current data exchange mechanisms will in

principle allow them to flow from the receiving database to the others (2) each group will have to compare *all* published papers against their stack of submissions awaiting citation completion - not just papers from the journals they scan. In view of these and other points about a geographical split, it was agreed that each group should go away and think about what the implementation of such a scheme would mean in terms of changes to their current processing mechanisms and about what additional information, if any, they would need from the other databases. This should be complete by mid-August, at which time we will compare notes and decide on the next steps.

Implementation will be a 3 stage process: (1) each group will begin to process ALL submissions they receive, including those relating to a journal scanned by another group (that is, the submission will no longer be forwarded to this other group); (2) we would invite NAR and perhaps one other (GenBank?) journal to try out a scheme in which authors are explicitly told that they may send submissions to any of the three nucleotide sequences databases; and (3) we will invite other journals to do the same.

Each group gave a progress report on the state of their journal interactions. A growing number of GenBank and EMBL journals are operating a "pseudo-mandatory" submission scheme in which they strongly state that authors must submit their data to the database, although in the end this isn't enforced (i.e., publication will still take place), and several journals in addition to NAR are either operating a true mandatory scheme or planning to begin doing so.

In view of the previous discussion about moving away from a journal split for handling data submissions, we discussed whether we should continue the "one database <-> one journal" model for interacting with journal editors; the decision was that we should.

7. Features table implementation schedule by P. Gilna

The discussion opened with each group reporting on the present status of implementation of the new features table.

GenBank reported that the new table was embodied in the flat-file report being written from the relational database and was planned for release in March 1990 (Rel. 63.0). GenBank did not envisage that annotators would use the FT directly, but rather through use of the Annotator's workbench. The new FT has also been implemented in the input to Authorin and in the quasi-GenBank format output provided by the program. The submission output of the program, an ASCII transaction protocol also embodies the FT. Such features will be manually translated from submissions recieved prior to the March 1990 release.

DDBJ reported that they were waiting for the GenBank implementation of the relational schema and would not be doing any work of their own accord until such time as the schema and supporting software were installed at Mishima. They expressed an urgent need to port to the new FT as soon as possible, as they did not wish to have to retrain (old -> new format) the cadre of new annotators they expected to hire in the near future. DDBJ further expressed tha intention of using Authorin as their primary data entry and annotation tool, planning to store all data as transactions and passing the data to the GenBank of EMBL in that format. They requested that they be allowed to begin to do so immediately. As EMBL would be unable to handle the transactions from Authorin, they requested that DDBJ wait until Genbank at least could handle the transactions and pass them in turn to EMBL.

EMBL stated that they were delayed in their incorporation of the new FT into their relational design and report generator, and that they would be some months behind GenBank in incorporating the FT in their releases, but that they were committed to it's incorporation. EMBL expressed concern at the programming effort needed to translate

GenBank new FT flat-files into EMBL old FT flat files in the interim period. GenBank acknowledged this and although not part of their development plans offered to investigate the production of the code necessary to generate the flat-file in the old FT format if requested by EMBL.

8. New line type for "data source" by C. Burks

As an extension of the discussion of tracking data citations and sources (see Topic #7 above), we discussed the possibility of a new, additional linetype or field (in the context of the flat file formats) for tracking the source of data.

In essence, the current REFERENCE block (as in GenBank) would be retained as a citation (that is, a pointer to a germane article) in the "classical" sense found in citations in journal articles ; but any current data items associated with the REFERENCE block that delineate the source of the data would be teased out, and combined with new data items, to give us greater flexibility in describing and tracking data sources.

With the unpropitious mid-stream status of conversion to the RDBMS at all three sites, introduction of a new linetype and/or redefinition of old linetypes was considered undesirable at this time. But we resolved to begin developing proposals that could be exchanged, discussed, and implemented as soon after the restructuring as possible.

A likely minimal set of data items to be tracked for data_source was determined to be: incoming receipt date; original port-of-entry among the three current "official" collaborators (e.g., "EMBL"); point of origin of data (e.g., "GenBank staff" or "J. Messing"); and medium of submission/acquisition from point of origin (e.g., "electronic mail".

9. CD-ROM progress and joint use by EMBL, GenBank, PIR by David Benton

David Benton gave an overview of the proposed CD ROM format file structures and distributed the most recent draft (dated 1 June) of the format proposal. GenBank is currently developing the software required to produce the database formatted according to the proposal. While the design of this software is still too preliminary to allow a precise estimate of its size to be made, it seems likely that an initial release could be made in spring 1990. Initial releases will include GenBank in its current flat-file and floppy-disk formats. GenBank will also distribute sample database access programs based on software developed to test the file formats and contents. The programming staff is currently investigating the possibility of providing the low-level database access functions in a callable library.

Graham Cameron reported on the EMBL experience with Phillips-DuPont in producing a CD ROM for distribution. The EMBL Data Library Nucleotide Sequence Database and the Swiss-Prot protein sequence database in the current flat file distribution formats, in a format suitable to the Circle Information Systems retrieval software, and in the NBRF format, and the aforementioned CIS retrieval software are distributed on the EMBL CD ROM. A first successful pressing (800 copies) was expected momentarily. As soon as possible, the nucleotide sequence database and Swiss-Prot in the common CD ROM format will be added to the EMBL CD ROM.

DNA Databases and Genome Projects

6/23/89		
19:00 - 21:00	Banquet at the Mishima Plaza hotel A welcome message	Ei Matsunaga Director of the NIG
6/24/89	At the National Institute of Genetics	
9:30 - 9:35	Opening	Akira Ishihama Chairman of the DDBJ advisor committee, NIG
9:35 - 10:05	Funding for genome projects in the U.S.A.	Jane Peterson GenBank grant officer NIGMS, NIH
10:05 - 10:45	Molecular biological databases	Christian Burks Co-principal investigator of the GenBank project LANL
(Break)		
11:00 - 11:40	GenBank project	David Benton GenBank Manager IntelliGenetics
11:40 - 12:20	Shifting the paradigm for data flow into GenBank	Paul Gilna GenBank, LANL
(Lunch)		
1:50 - 2:30	Restructuring of the GenBank database	Michael Cinkosky GenBank, LANL
2:30 - 3:10	Activity at the EMBL Data Library and Restructuring of the EMBL database	David Hazledine EMBL Data Library
(Break)		
3:40 - 4:20	On the sequencing of the whole Escherichia coli genome	Katsumi Isono Kobe Univ.
4:20 - 5:00	From mapping to sequencing of human genome	Eiichi Soeda Inst. Phys. Chem.
5:00 - 5:40	Computational approaches in the genome projects	Minoru Kanehisa Kyoto Univ.
5:40 - 5:45	Ending remarks	Takashi Seno Head of the DNA Research Center

Discussios:

N. Aiba, N. Asada, H. Asaine, Y. Fukumaki, T. Furuichi, T. Inoue, A. Ito, T. Koide, H. Kojima, A, Konagaya, Y. Kubota, Y. Maeda, A. Nakata, K. Nagai, K. Negishi, Y. Nishida, Y. Ogihara, K. Ohori, T. Ooi, H. Sakai, J. Sekizawa, Y. Seto, H. Tanaka, T. Tanaka, H. To, K. Tsukamoto, T. Tsuzuki, Y. Watanabe, K. Wakabayashi, H. Yamada, K. Yamakawa, Y. Yonezawa, K. Yoshida

国立遺伝学研究所

研究集会

「ＤＮＡデータベースとゲノム解析」

1989年 6月24日

参加者名簿　（順不同）

磯野　克己	（神戸大学理学部）	添田　栄一	（理化学研究所ジーンバンク）
金久　實	（京都大学化学研究所）	伊藤　彬	（癌研究会癌研究所）
大井　龍夫	（京都女子大学家政学部）	塚本　喜久雄	（千葉大学薬学部）
中田　篤男	（大阪大学微生物研究所）	関沢　純	（国立衛生試験所化学物質情報部）
酒井　裕	（京都大学　農学部）	萩原　保成	（横浜市立大学木原生物研究所）
山川　和弘	（東洋紡績）	服巻　保幸	（九州大学遺伝情報実験施設）
米澤　保雄	（帝人）	根岸　和雄	（岡山大学遺伝子実験施設
渡辺　良成	（福井医科大学）	續　伯彦	（愛知学院大学電子計算機センター）
西田　康伸	（中京大学教養部）	小嶋　洋之	（工業技術院大阪工業技術試験所）
田中　龍夫	（山形大学医学部）	愛場　信康	（富山医科薬科大学第３内科）
藤　博幸	（蛋白工学研究所）	古市　貞一	（基礎生物学研究所）
井上　貴文	（基礎生物学研究所）	小出　剛	（基礎生物学研究所）
浅田　伸彦	（岡山理科大学理学部）	前田　好美	（東京大学応用微生物研究所）
瀬戸　保彦	（蛋白質研究奨励会）	窪田　綏	（ノヴァ）
若林　一彦	（山梨医科大学）	永井　啓一	（日立製作所）
朝稲　宏文	（三井情報開発）	小長谷明彦	（日本電気 C&Cシステム研究所）
田中　秀俊	（ＩＣＯＴ）	吉田　かおる	（ＩＣＯＴ）
大堀　兼男	（浜松医科大学）	山田　久義	（東洋醸造技術情報センター）

FLAT DB E-Mail Network Server

E-mail address for the server: netserv@ddbj.nig.ac.jp
E-mail address for inquiries: sanzo.miyazawa@ddbj.nig.ac.jp

Following commands are available.

man subjects
Output manuals.

scandir db-name [options] 'keyword[|keyword...]'
Scan directory files of the "db-name" database to find "keywords" and output entry names and their definitions; keywords should be expressed in the regular expression.
Options:
-i case insensitive

scanjou db-name 'journal'
Scan journal index files of the "db-name" database to find specified journals and output journal names and corresponding entry names.
Journal names in the command line are not case sensitive.

scanaut db-name 'Last-name,[First.Middle-Initial.]'
Scan author index files of the "db-name" database to find specified author names and output author names and corresponding entry names with their definitions.
Author names in the command line are not case sensitive.

scanacc db-name '#acc'
Scan accession number index files of the "db-name" database to find specified accession numbers and output corresponding entry names with their definitions.

scandb db-name [-1] [-o] {['entry'...]|[-a '#acc'...]}
Scan the "db-name" database to find specified entries or accession numbers and output those entries.
Options:
-1 'Entry' or '#acc' may specify multiple entries in the DB.
-o The order of arguments is not significant; the order of entries output may not be in the order specified in the command line.

db-name = ddbj | gb | embl | swiss | pir | prf

ddbj: DDBJ DNA database; it is included in the GenBank and EMBL DBs.
gb or genbank: GenBank DNA database
Regular release and new entries which are updated twice a day.
embl: EMBL DNA database
swiss: SwissProt protein database
pir: PIR protein database
prf: Protein Research Foundation peptide database

Command names and others are case sensitive, unless specified.
Output is limited to 600 lines.

Examples:

```
scandir gb 'oncogene'
scanjou gb 'J. Biochem.' '1989'
scanaut gb 'Miyazawa,S.'
scandb gb 'AGMERLTR1'
scandb gb -a 'M11391'

scandir gb 'oncogene' | scandb gb       # get entries of "oncogene"
```

ニュースレター、ソフトウエア配布及び
DNA、蛋白質データベースの配布に関する活動報告

DDBJ　　堀江　元乃／鈴木　成子

米国からGenBankデータベース、欧州からEMBLデータベースを磁気テープで取り寄せ、希望者に配布しました。その他蛋白質データベース（SwissProtデータベース）も希望者には配布しています。配布媒体は磁気テープのみで、配布形態は定期もしくは一次配布です。磁気テープの配布総数は1989年 450本（1987年 504本、1988年 693本）、フロッピーディスクの配布枚数は1989年 1296枚（1987年 724枚、1988年 572枚、GenBank 24枚／件、DDBJ 2枚／件）です。

今年度の配布実績の詳細は以下のとおりです。

	大学／研究所	営利企業	合　計
DDBJニュースレター			
No.6 配布数	792	153	845
No.7 配布数	416	77	493
（この他日本がん学会、生化学学会等でも配布）			
No.8 配布数	372	94	466
（この他日本がん学会、生化学学会、分子生物学会等でも配布）			
定期配布希望者数(88/02/01)	121 (155)	23 (29)	144 (184)
定期配布希望者数(88/11/30)	201 (231)	43 (51)	244 (282)
定期配布希望者数(90/01/19)	232 (273)	65 (75)	297 (348)
	（　）内は配布総部数		
計算機所外利用者数　1987年度	50	9	59
計算機所外利用者数　1988年度	64（更新 34）	11（更新　6）	75
計算機所外利用者数　1989年度	68（更新 38）	17（更新 11）	85
Kermitプログラム配布件数　1987年度	32	5	37
Kermitプログラム配布件数　1988年度	11	8	19
Kermitプログラム配布件数　1989年度	53	22	75
VT emulator 配布数　1987年度	11	1	12
VT emulator 配布数　1988年度	7	5	12

GenBank　　(90/01/19　現在)

		定期配布		一時配布		合　計	
版		大学	企業	大学	企業	大学	企業
40	86/02	1	0	0	0	1	0
42	86/05	6	4	1	1	7	5
44	86/08	13	10	3	1	16	11
48	87/02	10	10	2	1	12	11
50	87/05	19	11	4	2	23	13
54	87/12	22	11	3	1	25	12
55	88/03	21	11	0	5	21	16
56	88/06	20	9	1	2	21	11
57	88/09	12	8	0	3	12	11
58	88/12	29	16	1	3	30	19
59	89/03	19	12	0	0	19	12
60	89/06	21	15	2	4	23	19
61	89/09	19	16	1	2	20	18

GenBank　圧縮版　フロッピー　　(88/11/30　現在)

		定期配布		一時配布		合　計	
版		大学	企業	大学	企業	大学	企業
40	86/02	1	0	1	0	2	0
44	86/08	10	0	28	5	38	5
48	87/03	1	1	5	2	6	3
52	87/08	4	2	9	1	13	3

(配布中止)

EMBL　　(90/01/19　現在)

		定期配布		一時配布		合　計	
版		大学	企業	大学	企業	大学	企業
8	86/04	6	5	3	2	9	7
9	86/09	10	8	1	0	11	8
10	86/12	10	8	0	0	10	8
11	87/04	12	8	3	5	15	13
12	87/07	11	8	0	0	11	8
13	87/10	15	10	0	0	15	10
14	88/01	16	15	3	2	19	17
15	88/05	13	12	1	3	14	15
16	88/08	7	6	1	2	8	8
17	88/11	22	18	1	3	23	21
18	89/02	19	14	1	0	20	14
19	89/05	15	14	2	3	17	17
20	89/08	18	18	1	0	19	18
21	89/11	7	4	0	0	7	14

DDBJ　フロッピー　　(90/01/19　現在)

	定期配布		一時配布		合　計	
版	大学	企業	大学	企業	大学	企業
1	0	0	0	0	0	0
2	0	1	3	2	3	3
3	0	1	5	2	5	3
4	1	1	1	2	2	3

DDBJ　磁気テープ　　(90/01/19　現在)

	定期配布		一時配布		合　計	
版	大学	企業	大学	企業	大学	企業
4	2	1	0	2	2	3
5	0	0	1	1	1	1

SwissProt　　(90/01/19 現在)

版	定期配布 大学	定期配布 企業	一時配布 大学	一時配布 企業	合計 大学	合計 企業
7	3	5	0	1	3	6
11	3	5	0	1	3	6
12	0	0	0	0	0	0

NBRF　　(88/11/30　現在)

版		定期配布 大学	定期配布 企業	一時配布 大学	一時配布 企業	合計 大学	合計 企業
27	86/03	4	3	2	0	6	3
28	86/07	5	5	2	1	7	6
29	86/09	8	7	1	0	9	7
30	87/01	8	7	0	0	8	7
31	87/06	10	5	0	0	10	5
32	87/11	9	7	2	0	11	9
33	88/05	11	6	1	0	12	6

配布は JIPID に移行

NBRF　VAX/VMS版　　(88/11/30　現在)

版		定期配布 大学	定期配布 企業	一時配布 大学	一時配布 企業	合計 大学	合計 企業
28	86/07	0	1	0	0	0	1
29	86/09	2	2	1	3	3	5
30	87/01	2	2	2	3	4	5
31	87/06	2	4	0	1	2	5
32	88/11	4	4	0	0	4	4
33	88/05	5	5	0	3	5	8

配布は JIPID に移行

PIR　　(88/11/30 現在)

		定期配布		一時配布		合　計	
版		大学	企業	大学	企業	大学	企業
8	86/02	9	6	2	2	11	8
10	86/08	12	8	0	0	12	8
11	86/12	13	8	0	1	13	9
12	87/03	16	6	2	1	18	7
13	87/06	16	6	0	1	16	7
14	87/09	17	6	1	0	18	6
15	87/12	16	7	0	0	16	7
16	88/03	16	7	1	1	17	8
17	88/06	15	9	0	2	15	11
18	88/09	5	8	0	1	5	9

配布は JIPID に移行

PIR　VAX/VMS版　　(88/11/30　現在)

		定期配布		一時配布		合　計	
版		大学	企業	大学	企業	大学	企業
8	86/02	0	1	0	0	0	1
9	86/05	0	1	0	0	0	1
10	86/08	3	4	0	1	3	5
11	86/12	4	5	0	0	4	5
12	87/03	4	7	0	1	4	8
13	87/06	3	8	0	0	3	8
14	87/09	4	7	0	0	4	7
15	87/12	5	8	0	2	5	10
16	88/03	5	9	0	0	5	9
17	88/06	6	7	0	1	6	8
18	88/09	2	4	0	0	2	4

配布は JIPID に移行

Final Report: DDBJ/EMBL/GenBank Meeting

5-9 February 1990

Representatives of the staffs of the DNA Data Bank of Japan, the EMBL Data Library, and GenBank met in Taos, New Mexico (U.S.A.) to discuss ongoing collaborative issues. Though several topics were discussed, the primary topic of discussion was unification of the databases and, in particular, strategies for achieving functional equivalence in the near term.

The topics discussed and decisions reached are summarized below; in each case, the chairperson's name for any particular session is given in parentheses after the topic heading.

1.0 Unification Issues (P. Gilna)

The opening session of the meeting focused on the elaboration of the overall goal of the collaborative meeting. The need for improved data exchange was first placed in perspective: over the past seven years, EMBL and GenBank, while exchanging data in the flat file format, have not propagated entry updates to each other except for new citations associated with existing entries. It was estimated that to correct this situation at this point (i.e., to merge the databases, remove duplicate entries, and reconcile related but non-identical entries) would take on the order of 6 person-years to perform manually; and though it would go faster with more automation, staff time for development for the automatic tools would reduce the net gain of automation for this task considerably. Without a system in place to propagate updates, the outcome of this task will be outpaced by the growth in new unpropagated updates over the course of the project.

While it was mentioned that the present level of data exchange was already considered by some outsiders to be impressive in comparison to other collaborating databases, and that the community probably suffered more from intra- than inter-database inconsistencies, it was nonetheless agreed that the attention of the community on the collaboration was now particularly focused on incomplete data exchange, or more accurately, the lack of equivalence between the three databases.

While the application of the existing GenBank Transaction Protocol, or NCBI's proposal to adopt the ISO ASN.1 standard in defining entities were discussed, it was concluded that these were going to be long term approaches to implement, and that an immediate, interim solution was required. Furthermore, it was acknowledged that complete identity could not be achieved -- at least in the near term -- in the context of independent, autonomous databases. As a practical alternative, the concept of *functional equivalence* was examined, with the pragmatic goal of designing a system where, for a specified set of common elements, a scientist need only access one database (GenBank, EMBL, or DDBJ) to obtain all of the data associated with these elements.

The following goals were set for this meeting on the topic of functional equivalence:

- A draft definition of functional equivalence would be agreed upon.
- A list of common entities (and their attributes) to be exchanged and updated would be drawn up at this meeting.
- Possible mechanisms for automated exchange of the agreed-upon data elements would be evaluated, one would be chosen, and documentation outlining the implementation would be produced at this meeting.
- Schedules for the implementation of the agreed-upon mechanism(s) would be specified.

1.1 Update on DBMS Restructuring

1.11 GenBank Report (M. Cinkosky)

GenBank reported that they will be producing Release 63 from the relational database, and will be introducing the new feature table format. This will be accomplished by translating the normal release into the relational version and then producing the distribution version as a report.

The Annotator's Workbench has been in alpha testing with a small number of annotators and data entry staff for several months and should be ready for regular use by early May. Until this time, the annotation and data entry staffs will continue to use the current tools, producing entries in the flatfile format, which will then be translated into the relational database. Once the Annotator's Workbench is ready, the GenBank staff will switch over to working directly on the data in relational form.

Work has also been completed on a general-purpose transaction parser which is able to be automatically reconfigured based on the contents of the entity definition file contained in the Transaction Protocol Manual. This parser reads transactions in ASCII form into memory, where they can be executed by whatever software is appropriate. This ability to be easily reconfigured allows the transaction protocol to be used in any other context, with any entity definitions, without requiring programming to parse the transactions.

GenBank also reported success in writing the software required to parse the free-text fields (i.e., source and feature descriptions) from the flatfile format into atomic data items. This software will be run after the conversion to the relational system has been completed.

1.12 EMBL Report (D. Hazledine)

The installation of EMBL's nucleotide sequence database in its RDBMS schema, and the construction of interfaces between the schema and the existing data entry tools, was completed in autumn 1989. The first quarterly release to be distributed after the completion of the RDBMS conversion was Release 21 in November 1989.

EMBL will be working on the design and implementation of new data entry tools which will take full advantage of the RDBMS over the next few months, and plan to make the first production release of these tools by Autumn 1990.

1.13 DDBJ Report (S. Miyazawa)

Building their own RDBMS would require them to develop a significant amount of software for data entry and also for data exchange between databanks. This is too great a task for them, given their present level of funding and resources. So, DDBJ has decided to use the same RDBMS, Sybase, as GenBank uses; DDBJ will become a satellite of the GenBank RDBMS, with necessary software provided by GenBank. Use of the same RDBMS will allow us to build a true distributed database between DDBJ and GenBank.

Sybase was installed December 1989. Software developed by GenBank will be installed in March 1990.

1.2 Common Exchange Protocol (D. Benton)

A common transaction format, a common virtual schema (in fact, a common entity-relationship model of the databases), a list of allowable transactions, and the rules governing transaction processing were agreed to in draft form. These are included as a draft document in Appendix 1. The EMBL Data Library will maintain this document over the course of future revisions.

This protocol will be made final by 15 March 1990.

1.3 Implementation Plan (M. Cinkosky)

GenBank indicated that the implementation of the new data exchange format posed no serious problems for them. Because all updates to the relational version of GenBank are passed through the Data Access Library, which stores all successful transactions for forwarding on to satellite and peer databases, it is a simple matter to convert the appropriate transactions into the new format. They indicated that the actual implementation should require no more than two months, but because of the other changes being made on the database, they will delay the start of this task for about four months.

Because DDBJ will be using GenBank software, their implementation schedule is tied to the schedule at LANL.

EMBL indicated that they would prefer to wait to implement the new exchange mechanism until after they have completed their new annotation tools. This means that they expect to be able to begin this work in approximately six months. At this point, EMBL anticipates that this implementation should not present them with difficulties.

Based on these estimates, 1 March 1991 was chosen as the target date by which the new system would be in place at all sites.

2.0 Data Distribution

2.1. EMBL Report (D. Hazledine)

EMBL reported the following data distribution systems:

Copies of the entire database on magnetic tape continue to be the main way the data are distributed. In 1989 a number of new tape formats were supported. In April 1989 the Data Library produced a prototype release on CD-ROM, and since then regular releases have been distributed on this medium. The disk contains the nucleotide sequence database and SWISS-PROT and software to enable MS-DOS computers to search both databases. A common CDROM database format is being developed collaboratively and is being finalized by IntelliGenetics. EMBL will include data in this format as soon as its final specification is circulated.

The EMBL file server is now going into its third year of operation. The Data Library offers a continuously updated nucleotide sequence data collection via this facility. Also other databases and software are offered. At present the file server is processing between about 1000 requests per month. EMBnet was established in 1988 as a way of connecting molecular biological research centers throughout Europe for the purposes of data distribution, computer conferencing, and other information services. By February 1990 nearly all EMBL member states had identified a national node and nine nodes were already operational. The Data Library updates satellite copies of the nucleotide sequence database at the EMBnet nodes daily. It is intended that

national nodes should offer data distribution and analysis services within their country, and indeed a number are already doing so.

2.2 DDBJ Report (S. Miyazawa)

In December 1989, an e-mail network database server was made available at the DDBJ computer system through Junet, Usenet, Internet, and Bitnet. Available databases include PIR, SwissProt and PRF as well as DDBJ, EMBL, and GenBank databases. The EMBL and GenBank databases are updated twice a day by including new entries which are daily sent by e-mail from those databanks. The DDBJ database is updated as well. Functions provided by the server include keyword search on definition lines, search by author name, journal name and accession number, and entry retrieval by specifying entry names or accession numbers.

The DDBJ on-line service has been provided free to anyone since 1987.

2.3 GenBank Report (D. Benton)

Genbank data continue to be distributed four times per year on magnetic tape (9-track, TK-50, and QIC-24 media) in three formats (unlabelled fixed-record length files, VAX/VMS Backup Savesets, and UNIX tarfiles). In addition, twice a year an abridged version of the data is distributed on floppy diskettes (in IBM-PC double- and high-density formats and Macintosh 800-kb format). GenBank has added significant new functionality to its on-line service since September, 1989. These enhancements, including daily updates of the on-line databases, are described in detail in Appendix 2. GenBank will distribute the database in the common GenBank/EMBL/PIR CD ROM file format (generated as a report on the relational database) beginning in April, 1990.

3.0 New Data Items for Tracking Data Source (C. Burks)

Following up on the topic presented in the report on the previous collaborative meeting, we discussed implementation of data items providing more detailed tracking of the reports the sequence data are drawn from. Accordingly, four data items were added to the new common exchange protocol.

4.0 Common Editorial Policy and Documents (G. Cameron)

Progress on detailed data exchange mechanisms has been excellent, but, as the annotators were quick to point out, the mechanisms alone do not ensure uniformity of the information entered into the database by the various groups. Annotation is interpretative in nature, and so there can be differences of opinion as to how to represent information in the databases. The agreed exchange mechanism ensures that we can exchange a common set of data fields. The values that we enter into these fields is a matter of editorial policy. In many important areas the editorial policies of the databases are similar or identical but some aspects of it are only loosely documented. It was considered desirable to set up mechanisms to promote convergence of editorial policy. These will include:

- more careful documentation of editorial policy;
- exchange of existing and new documents;
- giving more priority to editorial policies at future meetings.

It was noted that computer systems which would allow shared documents to be accessed and updated by all sites would facilitate the development of common policies on this and other issues. Mechanisms such as our file servers could support the existence of shared documents. These possibilities will be pursued further at the next collaborative meeting.

To get this effort off the ground, J. Hayden at LANL will arrange for the initial exchange of annotation/editorial documents by 1 April 1990.

5.0 Journal Interaction and Split (P. Gilna)

The databanks are committed to the concept of removing the restrictions currently placed on authors whereby they must submit data to whichever database is responsible for scanning the journal in which the data will be published. It is proposed that authors should be allowed to submit to their database of choice, regardless of the databank bearing the scanning responsibility. Thus, authors in Europe could choose to submit data associated with a "GenBank journal" article to EMBL, Japanese authors may submit to DDBJ, etc.

By far the greatest issue this reises is that of how to establish links between data submissions and the journal articles in which the data are eventually published. This is particularly difficult when articles do not cite accession numbers. Each data bank presently performs this cross-linking in part manually, in part using software tools; they are currently improving the algorithms for this process. Given the difficulties in performing this process in-house, the prospect of having to do this across databases has led to some concern about data entry redundancy in the absence of a robust linking mechanism.

Given that one approach to improving the linking algorithms would involve increasing the number of data items being compared (e.g., sequence fragments), we discussed developing these additional data items, either with our own data entry mechanisms or by taking advantage of possible outside sources. One example of the latter possibility discussed was the NCBI InterGene database (currently under development). As such data sources come on-line, we will evaluate their potential for use in this way.

We should encourage journals (and provide them the enabling documentation) to incorporate changes corresponding to the author-oriented submission policies in their *Instructions to Authors*. We will begin negotiating with them as soon as we have addressed the need for better linking algorithms (as discussed above) and documented our own policy, keeping in mind that the journals only revise their *Instructions to Authors* once or twice a year and with a prolonged lead time for changes.

DDBJ expressed the need to implement this protocol as soon as possible, as it seemed particularly sensible that Japanese authors should be allowed to submit their data to DDBJ.

P. Gilna at LANL will coordinate the effort to compare and evaluate linking algorithms and develop a definitive policy to support the author-oriented journal split. He and others from GenBank, DDBJ, and EMBL will meet briefly to continue discussions of this topic immediately prior to the IAB meeting in Mishima in March 1990.

The notion of the "journal split" -- the list we maintain explicitly stating which database is currently responsible for scanning each journal -- will continue to be important even after we have implemented an author-oriented submission policy. Firstly, it will continue to be necessary to scan journals; secondly, it is essential that only one of the databases enter into negotiations with each journal on submission policies. To avoid possible confusion in discussions with editors and publishers in the future, it was decided that -- as a default, assuming no overriding concerns from editors or publishers -- journals being added to the scan list for the first time would be handled by the database in whose continent the copyright resides (no changes will be made for journals already assigned).

6.0 Features Table Implementation (P. Kahn)

Each group expected to begin distributing data in the new feature table format according to the following schedule:

- GenBank: release 63 (March 1990)
- EMBL: release 24 (August 1990)
- DDBJ: presently aiming for release 7 (July 1990)
  [but the final decision will depend on making sufficient progress with moving the data into the RDBMS.]

7.0 Changes to Flat File Format (P. Kahn)

It was generally agreed that the following developments would be highly desirable, once all groups have completed their conversion to a RDBMS:

- The databases should distinguish between two types of sequences:
  (a) atomic sequences corresponding to individual scientific reports; (b) merges of these atomic sequences. Making this distinction will serve many purposes, such as making it possible to specify different merges of the same underlying sequences.
- The specification of merges should be done using a sequence manipulation language.

We discussed whether it would be desirable to distribute the merged sequences themselves, or only the instructions for performing the merges (while recognizing that each group should be free to decide independently).

GenBank raised the point that entry names (e.g., LOCUS, ID) are a troublesome data item because it is difficult to keep their meaning consistent. This problem was recognized by the other groups, yet at the same time we all realize that users generally favor the idea of retaining entry names because they are easier to remember than accession numbers. Since automatic conversion from GenBank to EMBL requires that EMBL understand GenBank's naming rules, any changes which GenBank considers will be discussed with the other collaborative groups in advance.

8.0 Authorin, Submission Forms (P. Kahn)

AuthorIn release 1.1 (IBM PC) will be available in March 1990, and version 2.0 (Macintosh) will be released in September 1990. In the near future IntelliGenetics will provide EMBL with postcards which they can include in their mailings to people who submit data; these cards can be used by researchers to request copies of Authorin from IntelliGenetics.

LANL will investigate the feasibility of developing software to parse the machine-readable data submission form, so that the processing of these forms could be made more automatic.

9.0 Release Dates in Association with Released Data (G. Cameron)

It has been true for some time that the data banks include limited information with data entries indicating when the data were initially released and subsequently revised. However, it is also true that our approaches have been both incomplete and inconsistent.

We resolved to consider revisions of current policy on time-stamp policy and documentation, with the goal of implementing a better approach by January 1991.

10.0 Next Meeting and Final Report (C. Burks)

The next collaborative meeting will be held January 21-25, 1991 (or second choice of January 28 - February 1, 1991) in or near Heidelberg, FRG. The EMBL Data Library will make arrangements for this meeting.

C. Burks will compile and consolidate the written reports from the individual session chairpersons, circulate the final report in draft form prior to the IAB meeting in March 1990, and -- if time permits -- provide the final draft for distribution at the IAB meeting.

Appendix 1: DDBJ/EMBL/GenBank Data Exchange Protocol

Appendix 2: Recent changes in the GenBank On-line Service

Recent changes in the GenBank® On-line Service

David Benton
GenBank / IntelliGenetics, 700 E. El Camino Real, Mountain View, California, USA 94040

ABSTRACT

The GenBank On-line Service provides access to the GenBank and EMBL nucleic acid sequence databases and to the Swiss-Prot and GenPept protein sequence databases. Users can query the databases by sequence similarity and annotation keywords and retrieve entries of interest. This access is available through e-mail servers, anonymous FTP, anonymous interactive login, and login to established, password-protected, individual accounts.

INTRODUCTION

In November, 1989, GenBank added substantial new functionality to its On-line Service. This article describes the current services provided by the GenBank On-line Service (GOS) and methods of obtaining access to these services. Currently, the on-line databases include the most recent quarterly releases of GenBank, EMBL, GenPept, and Swiss-Prot, and the data added to each of these (except Swiss-Prot) since their most recent releases (in the New Data databases). The GenPept database is a protein and peptide sequence database derived by automatically translating the protein coding regions of annotated GenBank entries. The three New Data databases are updated daily. Access to all GOS services is available to both commercial and non-commercial users at the same cost. On-line help is available for all aspects of this Service and user manuals are available from the GenBank On-line Service at the author's address.

INTERACTIVE ACCESS

Interactive access to the GOS databases is provided through the Telenet public data network, via remote login over the Internet, and by direct-dial telephone lines. At present, the IRX (Information Retrieval Experimental Workbench) program (1) is the primary interactive database retrieval program. Fifteen-minute IRX sessions are provided to users via anonymous logins over the Telenet public data network and (using telnet) over the Internet. Extended use is available by login to either of two classes of established user accounts. The GenBank contract requires that the cost of providing these accounts be recovered from the users. Complete information on costs and application forms are available from the GenBank On-line Service at the author's address or (415) 962-7364.

Class 0 Accounts

Anonymous users of the interactive system are provided with fifteen-minute complimentary sessions using the IRX retrieval program. With this program, entries in any of the on-line databases can be located by searching for a keyword or combination of keywords appearing in any of the fields of the entries' annotations. Located entries can be displayed on the terminal or downloaded to the user's computer with the Kermit file-transfer program. (The Kermit program is available for a wide variety of computers from numerous software bulletin boards, user groups, and from Columbia University. MS-DOS and Macintosh versions are available from GenBank on request.) New users of the IRX program should read the on-line introduction and users guide which can be displayed by answering 'Y' to the first question the program asks ("Do you want help?").

To log in to the GOS Class 0 account, one must have a supported terminal or a computer with software for emulating one of those terminals (see the list in Figure 1) and a modem capable of communicating at 300, 1200, or (preferably) 2400 baud. Instructions for and an example of dialing a local Telenet number to access the GenBank computer are shown in Figure 1. The Telenet customer service number is (800) 336-0437 and can be used to obtain the number of the nearest Telenet node. After completing the login procedure shown in Figure 1, the IRX database query program is immediately started.

Class 1 Accounts

To avoid the fifteen-minute limitation on Class 0 interactive sessions, users of the GOS may wish to establish accounts on the GOS computer. These accounts provide access to the GOS computer, 1 Mbyte of disk space for user files, access to the Unix utilities, IRX, and interactive and batch mode use of FASTA and TFASTA (a version of FASTA that compares peptide sequences with nucleic acid sequence databases by translating the database sequences

in up to six reading frames "on the fly"). Class 1 accounts also provide electronic mail access for contacting other users of the GOS and users of computers connected to the Internet and other computer networks. (Restrictions on the use of the Internet for commercial purposes apply to GOS users. Please read the application materials carefully.) Access to a wide variety of electronic bulletin boards is also provided. A listing of some of the bulletin boards (also known as newsgroups) is provided in Figure 2. Of special interest are the bionet.journals.contents newsgroup which provides tables of contents of several important journals on-line before publication, bionet.sci-resources which provides on-line copies of the NIH Guides to Grants and Contracts, and bionet.molbio.genome-program which provides access to the staff of and announcements from the National Center for Human Genome Research of the NIH and the U.S. Department of Energy's Genome Program.

Class 2 Accounts

For an additional fee, Class 2 users are provided with access to the IntelliGenetics Suite of sequence analysis programs and databases formatted for those programs. Additional databases (e.g., the PIR Protein Sequence Database, KeyBank™, and VectorBank™) are also available to Class 2 users. Class 2 users also have access to all the facilities available to Class 1 users.

E-MAIL SERVERS

In addition to providing interactive access, GenBank currently offers two electronic mail servers, one for sequence similarity searching and one for database entry retrieval. These are freely available to anyone who can send mail to an Internet address. The following networks have gateways to the Internet: BITNET, EARN, NETNORTH and JANET. Users of computers on these networks may need to change the format of the addresses given below to send the message through a forwarding gateway. Users should consult their computer system managers or administrators to determine the proper forwarding gateway and address form. Questions regarding the use of the e-mail servers (or other aspects of the GOS) may be addressed to: CONSULTANT@GENBANK.BIO.NET. (For users without nameservers, GENBANK.BIO.NET has the IP address: 134.172.1.160.)

FASTA Server

The GenBank FASTA Server receives mail messages containing a nucleic acid or protein query sequence and instructions for the search, performs a FASTA (2) sequence similarity search against the specified database, and returns the results by electronic mail.

To access the program, send an electronic mail message containing the formatted query sequence (as described below) to the following Internet address: SEARCH@GENBANK.BIO.NET. To receive instructions on using the FASTA Server, send a mail message to this address containing the word **HELP** as its only text.

Queries consist of a mail message with search parameters identifying the database to be searched, values related to the search and the query sequence to be used in the search. Figure 3 is an example of a mail message sent for a FASTA search. The mail message has two mandatory lines, three optional lines, a line identifying the query sequence as described below, and lines containing the query sequence. These lines start with the keywords shown and appear in the body of the mail message in the following order:

DATALIB — This line specifies the database or database division to be searched. (mandatory). To search the entire EMBL, GenPept, or Swiss-Prot database enter **EMBL**, **GenPept**, or **Swiss-Prot**, respectively. The New Data databases can be searched by specifying **GenBank/new**, **EMBL/new**, or **GenPept/new**, respectively. GenBank can be searched in full or in part by specifying the division you wish to search. The choices are listed below and should be entered as shown:

GenBank/primate	GenBank/plant	GenBank/phage
GenBank/rodent	GenBank/organelle	GenBank/synthetic
GenBank/other_mammalian	GenBank/bacterial	GenBank/unannotated
GenBank/other_vertebrate	GenBank/structural_rna	GenBank/all
GenBank/invertebrate	GenBank/viral	GenBank/new

KTUP — This line identifies the k-tuple value which specifies the sensitivity of the search. Values range from 3 to 6 for nucleic acid searches (defaults to 4 if not entered) and from 1 to 2 for protein searches (defaults to 1 if not entered). Lower values result in more sensitive searches but require more time to complete. (optional)

SCORES	This line specifies the number of best-ranked sequences to be listed in the results. The default value is 100. (optional)
ALIGNMENTS	This line identifies the maximum number of best-ranked sequences to be aligned in the results. The default value is 20. (optional)
BEGIN	Contains only the keyword BEGIN. (mandatory)

The remainder of the message contains the query sequence identification line and the sequence data. In Figure 3 the query sequence and identifier are shown in "Pearson format". The query sequence identification line is a mandatory comment line which begins with a greater-than sign (">") followed by the name of the sequence, a space, and an optional note about the sequence. The sequence data begin on the next line. Sequences in IntelliGenetics/BIONET format (3) are also acceptable.

The query message format must be followed precisely, but note that either upper or lower case letters may be used. Each line of the message must contain less than 80 characters. Longer lines will be truncated. The message text begins with the keyword DATALIB and should not contain blank lines. The message should contain only one query sequence.

When a query message is received, it is placed in a batch queue, processed in the order received, and the results delivered by return mail. The status of a job being processed can be determined by sending a mail message to the SEARCH address above containing only the word **QUEUE**. No individual is permitted to have more than one search waiting in the queue at a time. If a user sends a second query message before his first request has been processed, the initial search will be cancelled and replaced by the search submitted second.

Entry Server

E-mail access to sequence database entries is provided for three primary reasons: 1) to enable users of the FASTA Server to retrieve entries identified by sequence similarity searches; 2) to enable users of the Class 0 interactive system described above, who access it by network remote login (e.g., telnet) to retrieve copies of entries of interest; and 3) to enable readers of journals that identify published sequences by accession number to retrieve computer-readable versions of those sequences. To retrieve a database entry, send a mail message containing only the entry name or the accession number (not both) to the address: RETRIEVE@GENBANK.BIO.NET. The on-line databases are searched and the entries (if any) that correspond to the supplied entry name or accession number are returned by electronic mail. To receive instructions on using the Entry Server, send a mail message to the RETRIEVE address (above) containing only the word **HELP**. Note that, due to the order in which the databases are searched, if both GenBank and EMBL data banks contain entries with the same primary accession number (the usual case), a query on the accession number will return the GenBank version of the entry. If the EMBL-format version of the entry is required, it can be retrieved from the EMBL file server (4).

ANONYMOUS FTP

In addition to interactive access and electronic mail servers, GenBank also currently provides files for anonymous FTP (File Transfer Protocol), including GenBank and EMBL new data and contributed software. Each week, the new entries created in the GenBank database are collected into an update file. The file has a name of the form gb*MMDD*.seq, where *MM* is the number of the month and *DD* is the date of file creation. Likewise, new EMBL entries are collected into files with names of the form em*MMDD*.seq. The weekly update files are kept in the new data directories until they are superseded by a new quarterly release of the database.

Currently, the following directories are available for anonymous FTP:

pub/db/alu	Jurek Jurka's Alu sequence and alignment database
pub/db/embl-newdata	EMBL weekly update files
pub/db/gb-newdata	GenBank weekly update files
pub/db/gb-rel62	GenBank release 62.0
pub/db/gp-newdata	GenPept weekly update files
pub/db/gp-rel62	GenPept data derived from GenBank release 62.0
pub/db/seqanalref	Amos Bairoch's bibliography of sequence analysis literature
pub/dos	IBM PC compatible programs
pub/mac	Macintosh programs
pub/vms	VAX/VMS programs

The weekly update files are available as standard ASCII files or as compressed ASCII files. The compressed files are about one-third the size of the standard files. They can be distinguished by the .Z suffix and can be uncompressed

after transfer with the standard Unix uncompress utility. In addition to the weekly incremental updates, a cumulative update file, updated daily and containing all entries which have entered the database or been revised since the previous release, is maintained in each of the new data directories. These files, which are provided in compressed form only, are named gbseq.all.Z, gpseq.all.Z, and emseq.all.Z. The current GenBank and GenPept release files (in gb-rel*NN* and gp-rel*NN*, respectively, where *NN* is the release number) are provided in compressed form only. The software in the dos, mac, and vms directories is not supported by the GenBank On-line Service.

To access any of the directories available for anonymous FTP, one should use the FTP protocol to connect to GENBANK.BIO.NET, using **anonymous** as the Username and one's surname as the Password. The files in these directories are also available for downloading to users who access the GOS computer via Telenet or direct-dial. High-speed modems are provided for the direct-dial lines to facilitate file downloading.

ENVIRONMENT

All components of the GOS operate on a Solbourne Series 4/802 multiprocessor superminicomputer capable of 80 million instructions per second (MIPS). This computer is currently supplied with 64 Mbytes main memory, 3.5 Gbytes of magnetic disk storage, 2 Gbytes of on-line archival tape storage, a 9-track tape drive, and optical disk drives. This computer is dedicated to the GOS; other GenBank work, including GOS development work, is performed on other computers. The GOS computer uses the Unix operating system (Sun OS 4, a dialect of bsd 4.3). The GOS computer is connected to the Bay Area Regional Research Network (BARRNet) by a T1 (1.54 Mbit/sec) communication channel. BARRNet is the local branch of the NSFNet.

CONCLUSION AND FUTURE DIRECTIONS

Recent improvements in the GenBank On-line Service now make it possible for users to access daily updates of the databases via electronic mail or through easy-to-use and powerful interactive programs. Those who maintain local copies of the database can now update these copies by anonymous FTP on a weekly basis. In addition, similarity searching and analytical software are now available to users of the GOS. Within the next year, we plan to add access to the relational database version of GenBank. This addition will make possible much more complex queries than can be posed using existing database query programs.

ACKNOWLEDGEMENTS

The author gratefully acknowledges the major contributions of the following to the work described in this article: Stephen Barnhouse, Michael Cinkosky, David Horner, Rand Huntzinger, David Kristofferson, Will Nelson, Eliot Lear, Julie Ryals, Peter Stoehr, Spencer Yeh, and Katherine Yudin. The author thanks Doug Brutlag and John Moore for helpful comments on the manuscript. Provision of the IRX program by Dr. Dennis Benson and the National Library of Medicine and the FASTA program by Dr. William Pearson are also gratefully acknowledged. This work is funded by NIH Contract N01-GM-7-2110. GenBank is the registered trademark of the National Institutes of Health. KeyBank and VectorBank are trademarks of IntelliGenetics Incorporated. PIR is a registered trademark of the National Biomedical Research Foundation.

REFERENCES

1. Harman, D., Benson, D., Fitzpatrick, L., Huntzinger, R., and Goldstein, C. (1988) In Proceedings of RIAO 88 Conference *User-Oriented Content-Based Text and Image Handling*, Cambridge, MA., pp. 840-848.
2. Pearson, W.R. and Lipman, D.J. (1988) Proc. Natl. Acad. Sci. U.S.A. **85**, 2444-2448.
3. Brutlag, D.L., Clayton, J., Friedland, P., and Kedes, L.H. (1982) Nucleic Acids Res. **10**, 279-294.
4. Stoehr, P.J. and Omond, R.A. (1989) Nucleic Acids Res. **17**, 6763-6764.

FIGURES

Figure 1. Typical Class 0 GOS Telenet login procedure. In the example above, the user types **bold** characters exactly as shown, types user-specific information similar to that shown in ***bold-italic***, and sees the messages shown in `typewriter` font. The symbol CR indicates that the key labelled "Return", "RET", or "Enter" should be pressed. Comments and instructions are shown in *italic*.

INTERNET BBOARD Name	USENET Newsgroup Name
AGEING	bionet.molbio.ageing
AGROFORESTRY	bionet.agroforestry
BIONEWS	bionet.general
BIOTECH	bionet.technology.general
BIO-CONVERSION	bionet.technology.conversion
BIO-JOURNALS	bionet.journals.contents
BIO-MATRIX	bionet.molbio.bio-matrix
BIO-SOFTWARE	bionet.software
EMBL-DATABANK	bionet.molbio.embldatabank
EMPLOYMENT	bionet.jobs
GENBANK-BB	bionet.molbio.genbank
GENOMIC-ORGANIZATION	bionet.molbio.gene-org
HUMAN-GENOME-PROGRAM	bionet.molbio.genome-program
METHODS-AND-REAGENTS	bionet.molbio.methds-reagnts
MOLECULAR-EVOLUTION	bionet.molbio.evolution
PIR	bionet.molbio.pir
POPULATION-BIOLOGY	bionet.population-bio
PROTEIN-ANALYSIS	bionet.molbio.proteins
RESEARCH-NEWS	bionet.molbio.news
SCIENCE-RESOURCES	bionet.sci-resources
SWISS-PROT	bionet.molbio.swiss-prot

Figure 2. International BIOSCI network bulletin boards available on the GenBank On-line Service. Internet and Usenet names of each bulletin board are shown.

```
From:  drbob@someaddress.somewhere.edu Tue Jun 14 21:36:38 1988
Date:  14 Jun 1988 2129:02-PDT
To:    SEARCH@GENBANK.BIO.NET
Subject:

DATALIB GenBank/other_mammalian
KTUP  4
SCORES 100
ALIGNMENTS 20
BEGIN
>BOVPRL GenBank entry BOVPRL from gbmam file.907 nucleotides.
tgcttggctgaggagccataggacgagagcttcctggtgaagtgtgtttcttgaaatcat
caccaccatggacagcaaa
```

Figure 3. Example FASTA Server e-mail query message. The first four lines are a mail header that is automatically created by the mailer program. The Subject may be left blank (it is ignored by the server).

DDBJ計算機 広域ネットワークに接続される：
東京大学、名古屋大学、京都大学、大阪大学、九州大学の
各計算機センターの UNIX 計算機からネットワークを経て
DDBJ 計算機の login が可能となる。

1990年 5月 3日
宮澤、DNA Data Bank of Japan

DDBJ は今後増大が予測されるDDBJ/EMBL/GenBank 間でのデータ交換、データベースの共同構築に対処するため 2 年前より米国の広域ネットワークへの接続を計画してきましたが、1990年 3月 28日 DDBJ Ethernet Local Area Network は三島ー東京間の専用線（64 kbps）を用い東京大学理学部の計算機に接続されることにより広域ネットワーク（Japan Academic Internet）に接続[1]されましたので報告します。

Japan Academic Internetは、研究上国外のネットワークへの接続を必要とする諸機関（国際理学ネットワーク[2]）、広域ネットワーク構築研究を目的とするグループ（WIDE プロジェクト[3]、文部省科学研究費グループ）により構築されボランテイアーの手で維持されています。このネットワークは、現在もっとも普及している TCP/IP 手順を用いたネットワークで 高速な光ケーブル FDDI(100Mb/s), もっとも普及している Ethernet Thick/Thin cable (10Base-10/5), 電話線 (Twisted cable: max. 100m) (10Base-T)等の媒体を用いた 10 Megabits/sec のローカルネットワークをサポートし、また専用デイジタル回線（192kbps, 64kbs, 48kbps, 19.2kbps, 9.6kbps ...）、ISDN 公衆回線（192kbs, 64kbps）を用いた広域ネットワークにおいて X.25 手順やその他の point-to-point 専用手順の上で実現されるものです。TCP/IP 手順は、N 1 手順と異なり世界中の多種類の計算機がサポートしており、現時点では国外のネットワークへの接続には欠かせません。

このネットワーク手順は米国国防省の研究プロジェクトとして推進され、米国ではInternet と呼ばれるネットワーク構築に使用されています。Internet は米国全土に広がり1986年の時点で約 2000ノード、現在では全米のほとんど全ての主要大学、研究所、企業が56 kbps から1.5 Mbpsの回線速度で接続されています。今度 DDBJ が接続された広域ネットワーク（Japan Academic Internet）もハワイ大学をへて米国 Internet, オーストラリア、ニュージーランドに接続されています。そこで DDBJ でもDDBJ/EMBL/GenBank 間でのデータ交換に4 月よりこれまで使用されていた公衆電話回線に代わりこのネットワークを使用しています。

一方国内でも 国際理学ネットワーク、WIDE プロジェクト、文部省科学研究費グループに属する計算機が Internet 接続された結果、それらの計算機が属する学内 LAN / 所内LANが相互に Internet 接続されました。それ故、日本全国の多数の計算機から ネットワークをへてDDBJ 計算機をアクセスすることが可能になりました。もちろんこれらの計算機の大部分は各研究室の所有であり一般利用できませんが、一般利用の許された計算機（東京大学、名古屋大学、京都大学、大阪大学、九州大学の各計算機センターに属する計算機）も少数含まれます。そこで DDBJ ではこれらの計算機の上に、DDBJ terminal server 用の特別のアカウントを用意しました。このアカウント用いれば、各計算機のアカウントを取

得しなくても DDBJ 計算機を login できます。（このアカウントの使用料は DDBJ で負担いたします。） DDBJ 計算機よりそちらの計算機の方が距離的に近い場合は、DDBJ 計算機へのアクセスのためにご利用ください。アクセスのための電話番号、手順は別紙を参考にしてください。しかしこのネットワーク接続はほとんどの場合計算機センターとは無関係ですし、またボランテイアーにより維持されていますので、疑問点等の問い合わせは各計算機センターではなく DDBJ にご連絡ください。くれぐれも各計算機センターに迷惑をかけないようお願いいたします。

Appendices:

1. TCP/IP 手順でサポートされるコマンドの例：

```
% telnet hostname.domains.network                リモートログイン機能
% ftp  hostname.domains.network                  ファイル転送
% mail user@hostname.domains.network             電子メール

UNIX 計算機間では以下のコマンドもサポートされる。
% rlogin hostname.domains.network                リモートログイン機能
% rcp hostname.domains.network：/tmp/file /tmp/file  ファイル転送
% rsh hostname.domains.network cat /tmp/file     リモートコマンド実行
% finger @hostname.domains.network               利用者の表示
% talk user@hostname.domains.network             talk
%
...
```

DDBJ 計算機のホスト名は ddbj.nig.ac.jp です。よってネットワーク上の計算機から DDBJ 計算機を login するには、

```
% telnet ddbj.nig.ac.jp        or % telnet 133.39.128.2
```

2. Japan Academic Internet のネットワーク図；別紙参照

3. DDBJ terminal servers リストとその使用例；別紙参照

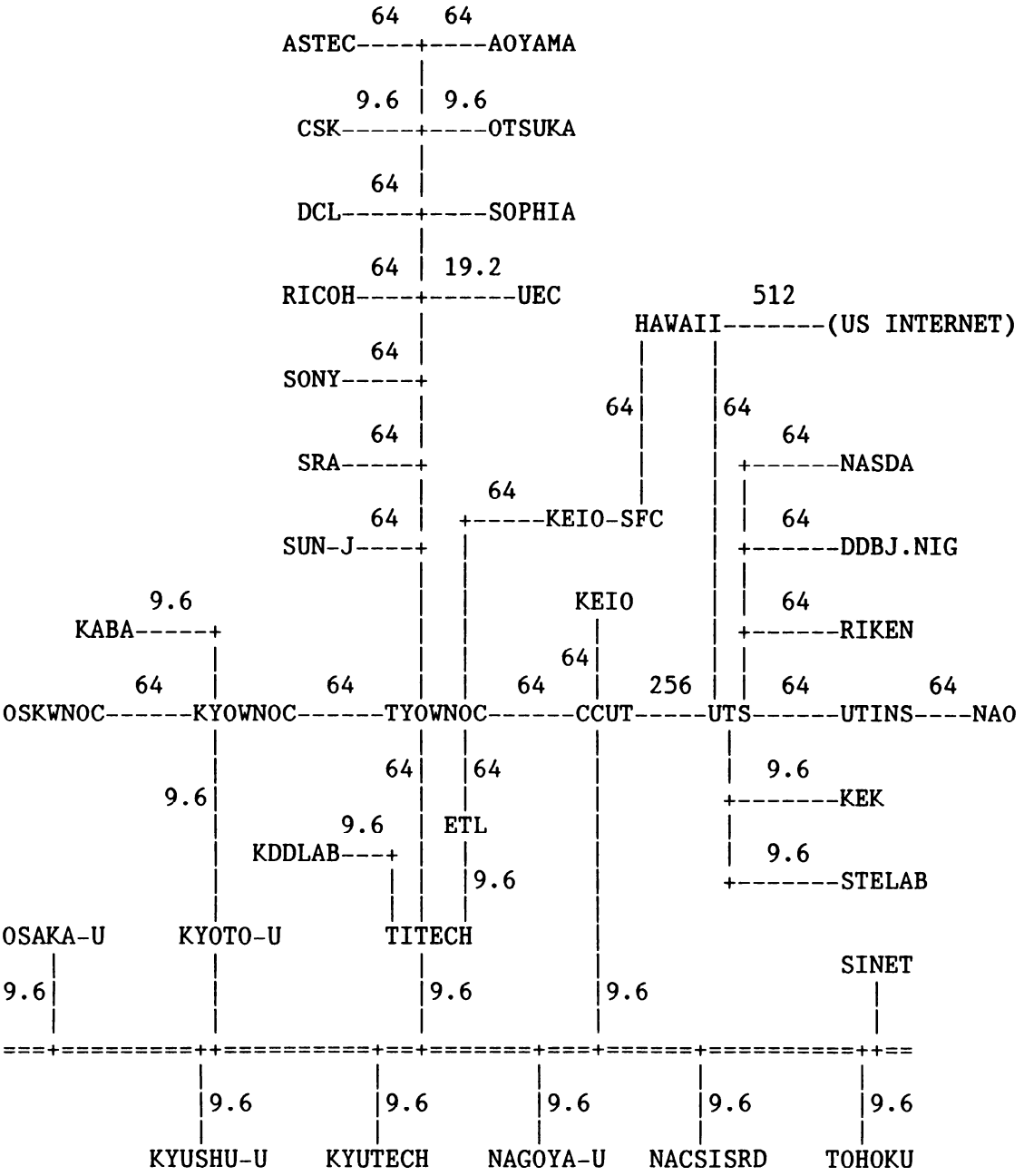
文献：

1. NEWS: GENETIC DATABANKS: New computer link for Japan, Nature vol. 344, p. 92, 1990
2. NEWS: COMPUTER NETWORKS: Japan ends its isolation, Nature vol. 340, p. 670, 1989
3. 村井純： WIDE プロジェクト、bit, vol. 22, pp. 118-126, 1990, 共立出版

*** JAPAN ACADEMIC INTERNET ***

JAPAN ACADEMIC INTERNET（仮称）を構成するネットワークプロジェクトは以下の３つである。

1. 科研費総合研究(A)　「わが国における大学ネットワークの相互接続に関する研究」
2. WIDEプロジェクト
3. 東京大学理学部 国際理学ネットワーク（TISN）



数字はボーレート Kbps
学情網の上のバーチャルサーキットの図示は省略した
ルーティング制限のあるリンクは図示していない
* Taken from ip-map written by Hiroaki Takada (hiro@is.s.u-tokyo.ac.jp)

略号

TISN （国際理学ネットワーク）
DDBJ.NIG = DNA Data Bank of Japan, 国立遺伝学研究所
RIKEN = 理化学研究所
KEK = 高エネルギー物理学研究所
NAO = 国立天文台
NASDA = 国立宇宙開発事業団
STELAB = 名古屋大学空電研究所
UTS = 東京大学理学部（TISN-NOC）
UTINS = 東京大学原子核研究所（on 田無）

WIDE
AOYAMA = 青山大学
CCUT = 東京大学大型計算機センター
ETL = 電子技術総合研究所
KEIO = 慶応大学
KEIO-SFC = 慶応大学湘南藤沢キャンパス
KYOTO-U = 京都大学
KYOWNOC = WIDE京都ネットワークオペレーションセンター（ASTEM）
KYUSHU-U = 九州大学
NAGOYA-U = 名古屋大学
OSAKA-U = 大阪大学
OSKWNOC = WIDE大阪ネットワークオペレーションセンター
OTSUKA = 筑波大学大塚キャンパス
SOPHIA = 上智大学
TITECH = 東京工業大学
TYOWNOC = WIDE東京ネットワークオペレーションセンター
UEC = 電気通信大学
その他 略

科研費総合研究(A) 「わが国における大学ネットワークの相互接続に関する研究」
NACSISRD = 学術情報センター研究開発部
SINET = 学術情報センター ＳＩメイルゲートウェイ
CCUT = 東京大学大型計算機センター
KYOTO-U = 京都大学
KYUSHU-U = 九州大学
NAGOYA-U = 名古屋大学
OSAKA-U = 大阪大学
TITECH = 東京工業大学
TOHOKU = 東北大学

References:

1. 村井純：WIDE プロジェクト，bit, vol. 22, pp. 118-126, 1990, 共立出版

2. NEWS: COMPUTER NETWORKS: Japan ends its isolation, Nature vol. 340, p. 670, 1989

3. NEWS: GENETIC DATABANKS: New computer link for Japan, Nature vol. 344, p. 92, 1990

List of terminal servers at computers
from which DDBJ is accessible via Japanese Academic Internet.

DNA Data Bank of Japan
National Institute of Genetics

Organizations	Host name	bps	phone/others	(terminal server for DDBJ) loginname	password
Tokyo Univ. VAX 8600	tansei	64kbps	03-814-7271 7 bits even parity 2400,9600 bps	a88657	DDBJ-NIG
NIG Sun 4/390	ddbjs4s	64kbps	0559-75-6036,6037 DDX-P: (163-060)-522-5127 8 bit no parity 1200,2400 bps full hostname: ddbjs4s.ddbj.nig.ac.jp ip-address: 133.39.128.2 For remote-login at a site on the Internet % telnet ddbj.nig.ac.jp or % telnet 133.39.128.2	(direct login) ddbjnews	(no password)
Nagoya Univ. FACOM A600/UTS	nucc	9.6kbps	052-781-3442 7 bits even parity 2400 bps	a88657a	DDBJ-NIG
Kyuto Univ. FACOM M780/UTS	kuduts	9.6kbps	075-753-7478 8 bits no parity 2400 bps	a88657	DDBJ-NIG
Osaka Univ. work stations	ccsun01 ccews01 ccews08 ccews09	9.6kbps	06-876-3145 7bits even parity 1200,9600 bps	a88657a a88657a a88657a a88657a	DDBJ-NIG DDBJ-NIG DDBJ-NIG DDBJ-NIG
Kyushu Univ. FACOM M780/UTS	kyu-cc	9.6kbps	092-631-3278 8 bits no parity 2400 bps	a88657a	DDBJ-NIG

by Sanzo Miyazawa
sanzo.miyazawa@ddbj.nig.ac.jp

Remarks: Session Logs
---

1) Tokyo Univ.: 7 bits even parity

```
atdt038147271
CONNECT

login: a88657
Password:
Last login: Sat Apr 21 14:51:32 from ddbjs4s.ddbj.nig
Ultrix-32 V3.1 (Rev. 9) System #8: Thu Nov 30 17:39:06 GMT+9:00 1989
System will shutdown at Fri Apr 27 09:00:00 1990
System will start at    Fri Apr 27 17:00:00 1990
You have mail.
Good afternoon!


*** Welcome to the DDBJ terminal server; please make sure PARITY is NONE.

   Hosts to which you can connect:
     ddbj                DDBJ computer
     utvos3              HITAC M682/680 in the Tokyo Univ.

*** Type "logout" or a host name to connect: ddbj
Trying...
Connected to 133.39.128.2.
Escape character is '^]'.


SunOS UNIX (ddbjs4s)

login:
```

2) Nagoya Univ.: initially 7 bits even parity and then 8 bits no parity

Prompt	Type
	atdt052-781-3442
CONNECT	
	<CR>
enter class	33
class start	
	<CR>/<BREAK>s until prompt comes out
nucc	
(2400BAUD)	
login:	a88657a
password:	DDBJ-NIG

```
atdt052-781-3442
CONNECT


enter class
class start
x~xx
nucc

  Welcome to NUCC UTS


(9600 BAUD)
login: a88657a
Password:
########  uts news  ##################################################
#   5/1(Tue.) ha, nagoya-u anniversary ni atari, service ha,       #
#     okonai masen.                                                #
#   /usr1 no file ni yoyuu ga naitame, 5 gatsu ni narimashitara    #
#     keizoku shinsei no nai user-id no file wo delete shimasu.    #
##############################################  90/04/20  ############
Terminal Type: vt100


*** Welcome to the DDBJ terminal server; please make sure PARITY is NONE.

   Hosts to which you can connect:
     ddbj               DDBJ computer

*** Type "logout" or a host name to connect: ddbj
Trying...
Connected to 133.39.128.2.
Escape character is '^D'.


SunOS UNIX (ddbjs4s)

login:
```

3) Kyoto Univ.: 8 bits no parity

```
atdt075-753-7478
CONNECT
<CR>

kuduts

  Welcome to Data Processing Center, Kyoto University  UNIX (UTS/M V21L10).


login: a88657
Password:
Account id : a
       UTS charge           25 yen
       MSP charge           10 yen
     Total charge           35 yen
Estimated payment       10,000 yen
Current file space          24 Kbytes
File space limit        20,480 Kbytes
************************************************************
*                                                          *
*          Welcome to KUDUTS (M780 UTS UNIX)               *
*                                                          *
*      JUNET electric mail and news services available.    *
*      Type "rn" to read electric news.                    *
*      Our JUNET mail address is                           *
*           xxxxxx@kuduts.kudpc.kyoto-u.ac.jp              *
*                                                          *
*      File space limit has been raised to 20M bytes a user.  *
*                                                          *
************************************************************
You have mail.
Terminal Type (default vt100): vt100


*** Welcome to the DDBJ terminal server; please make sure PARITY is NONE.

   Hosts to which you can connect:
     ddbj               DDBJ computer

*** Type "logout" or a host name to connect: ddbj
Trying...
Connected to 133.39.128.2.
Escape character is '^D'.


SunOS UNIX (ddbjs4s)

login:
```

4) Osaka Univ.: initially 7 bits even parity and then 8 bits no parity

```
atdt06-876-3145
CONNECT
<CR>
*** COMPUTATION CENTER OSAKA UNIVERSITY ***


  CLASS        S Y S T E M          BPS

    1          ACOS & SX           1200
    2          ACOS & SX           2400
    3          ACOS & SX           9600
    5          WORKSTATION         1200
    6          WORKSTATION         9600
ENTER CLASS 5
```

4.2 BSD UNIX (ccsun01)

```
login: a88657a
Password:
Last login: Sat Apr 21 17:05:24 from 133.39.128.9
Sun UNIX 4.2 Release 3.4EXPORT (GENERIC) #1: Thu Apr 30 09:36:18 PDT 1987

          ************************************************
          ****  Computation Center, Osaka University  ****
          ************************************************


*** Welcome to the DDBJ terminal server; please make sure PARITY is NONE.

   Hosts to which you can connect:
     ddbj               DDBJ computer

*** Type "logout" or a host name to connect:
Trying 133.39.128.2 ...
Connected to 133.39.128.2.
Escape character is '^]'.
```

SunOS UNIX (ddbjs4s)

```
login:
```

5) Kyushu Univ.: 8 bits no parity
Connection between FACOM-M780/UTS and DDBJ may be unstable.

```
atdt092-631-3278
CONNECT
<CR>

kyu-cc

  Welcome to FUJITSU UTS


login: a88657a
Password:
***** statistics and charges ( a88657a  )                *****
**      total size of reserved files    =        4 KB      **
**      total charge                    =       99 yen     **
************************************************************


You have mail.
Terminal Type: vt100


*** Welcome to the DDBJ terminal server; please make sure PARITY is NONE.

   Hosts to which you can connect:
     ddbj               DDBJ computer

*** Type "logout" or a host name to connect: ddbj
Trying...
Connected to 133.39.128.2.
Escape character is '^D'.


SunOS UNIX (ddbjs4s)

login:
```

ＤＤＢＪ関連行事日程表

1987年 2月	DNA データバンク運営委員会
1987年 2月25-28日	EMBL/NIH Workshop "Future Database for Molecular Biology" （遺伝情報研究センター長　丸山　参加）
1987年 3月 1日	共同利用計算機の遺伝研内使用開始
1987年 5月	遺伝情報分析研究室　宮沢、GenBank 視察
1987年 7月16日	計算機接続用電話外線 5回線を敷設
1987年 7月	DDBJ DNAデータ Release 1.0 配布
1987年 8月 1日	計算機接続用DDX-パケット 1回線を敷設
1987年 8月	共同利用計算機の所外オンライン利用開始
1987年 8月	電子郵便開始 （GenBank,EMBLとの間の連絡を電子郵便に切り換える。）
1987年11月11日	（所外）DNAデータバンク運営委員会 （所内）DNAデータ研究利用委員会　共同開催
1987年11月11-20日	DDBJ-EMBL-GenBank annual meeting Intellienetics で開催。遺伝情報分析研究室　宮沢　参加
1988年 1月	DDBJ DNAデータ Release 2.0 配布
1988年 2月15-16日	第一回 International Advisory Committee on Biological Databases ワシントンD.C.にて開催
1988年 7月	DDBJ DNAデータ Release 3.0 配布
1988年 7月 4- 9日	DDBJ-EMBL-GenBank meeting: New Feature Tableの作成 EMBL で開催。遺伝情報分析研究室　宮沢　参加
1988年 9月 5-15日	DDBJ-EMBL-GenBank annual meeting EMBL で開催。遺伝情報分析研究室　宮沢、林田　参加
1988年12月 9日	（所外）DNAデータバンク運営委員会 （所内）DNAデータ研究利用委員会　共同開催
1989年 1月	DDBJ DNAデータ Release 4.0 配布
1989年 2月 3- 4日	第二回 International Advisory Committee on Biological Databases　EMBLにて開催
1989年 6月18- 22日	DDBJ-EMBL-GenBank meeting 遺伝研で開催。
1989年 6月23日	DNA databases and Genome Projects 研究集会 遺伝研で開催。
1989年 7月	DDBJ DNAデータ Release 5 配布
1990年 1月	DDBJ DNAデータ Release 6 配布
1990年 2月 5- 9日	DDBJ-EMBL-GenBank meeting GenBankで開催。
1990年 3月15- 17日	第三回 International Advisory Committee on Biological Databases　遺伝研にて開催

SEQUENCE DATA SUBMISSION FORM

This form solicits the information needed for a nucleotide or amino acid sequence database entry. By completing and returning it to us promptly you help us to enter your data in the database accurately and rapidly. These data will be shared among the following databases: **EMBL** Data Library (Heidelberg, W. Germany); **GenBank** (Los Alamos, NM, U.S.A. and Mountain View, CA, U.S.A), DNA Data Bank of Japan (**DDBJ**; Mishima, Japan); National Biomedical Research Foundation Protein Identification Resource (**NBRF-PIR**; Washington, D.C., U.S.A.); Martinsried Institute for Protein Sequence Data (**MIPS**; Martinsried, W. Germany) and International Protein Information Database in Japan (**JIPID**; Noda, Japan).

Please answer all questions which apply to your data. If you submit 2 or more non-contiguous sequences, copy and fill out this form for each additional sequence. When submitting nucleic acid sequences containing protein coding regions, please include a translation. Then send us **(1) this form, (2) a pre- or reprint of any manuscript** which pertains to these data, and **(3) your sequence data** (in one of the machine-readable formats described below) to:

DDBJ Submissions
Laboratory of Genetic Information Analysis
Center for Genetic Information Research
National Institute of Genetics
Mishima, Shizuoka 411, Japan
Phone: Japan (0559)75-0771 ext. 647, FAX: Japan (0559)75-6040
E-mail: *ddbjsub@ddbj.nig.ac.jp*
(*ddbj@ddbj.nig.ac.jp* for inquiries)

Please include in your submission any additional sequence data which is not reported in your manuscript but which has been reliably determined (for example, introns or flanking sequences).

When we receive this material we will assign the data an accession number, which serves as a reference that permanently identifies them in the database. We will inform you what accession number your data have been given and we recommend that you cite this number when referring to these data in publications.

If new data become available which would make the database entry more informative (e.g., function of the gene product or location of important sites within the sequence), or if you discover errors in the sequence, we urge you to contact us so that we can update your entry.

COMPUTER-READABLE DATA SUBMISSION FORM

A computer-readable form is available on the **distribution tapes** of the DDBJ, the GenBank, and the EMBL Data Library. **DDBJnews**(Japan 0559-75-6036), **BIONET**(Mountain View, CA, USA) and **SEQNET**(Cambridge, U.K.) also have copies. **We prefer you to use the computer-readable form** rather than this printed one. In this case, the form should be filled out with a text editor and sent via computer network or mailed to the address above. We will send you a computer-readable form upon request.

FORMATS FOR SUBMITTED DATA

We are happy to accept data submitted in any of the following **computer-readable formats: (1) Electronic file transfer:** files can be sent via computer network to: *ddbjsub@ddbj.nig.ac.jp*. These addresses can be reached via various gateways from ARPANET, BITNET, JUNET, JANET, etc. Ask your local network expert for help or phone us. **(2) Magnetic tapes:** 9-track only; 800, 1600 or 6250 bpi; ASCII (prefered) or EBCDIC character codes; unlabelled tape with fixed-length record and any blocksize (prefered), or standard labelled tape. **(3) Floppy disks:** we can read Macintosh diskettes, and 3-1/2" or 5-1/4" diskettes from MS-DOS systems.

Whatever format you choose, we would appreciate receiving the sequence data in a form which conforms as closely as possible to the following standards.

- Each sequence should include the names of the authors.
- Each distinct sequence should be listed separately using the same number of bases/residues per line. The length of each sequence in bases/ residues should be clearly indicated.
- Enumeration should begin with a "1" and continue in the direction 5' to 3' (or amino- to carboxy-terminus).
- Amino acid sequences should be listed using the one-letter code.
- The code for representing the sequence characters should conform to the IUPAC-IUB standards, which are described in: Nucl. Acids Res. 13: 3021- 3030 (1985) (for nucleic acids) and J. Biol. Chem. 243: 3557-3559 (1968) and Eur. J. Biochem 5: 151-153 (1968) (for amino acids). We prefer lower case characters for representing nucleic acids.

D1.5/4.89

I. GENERAL INFORMATION

Your last name	First name	Middle initials
Institution		
Address		
Computer mail address	Telex number	
Telephone	Telefax number	

On what medium and in what format are you sending us your sequence data? (see instructions on front page)

[ ] electronic mail

[ ] diskette: computer _____ operating system _____ editor _____

[ ] magnetic tape

record length _____ blocksize _____ label type _____

density [ ] 800 [ ] 1600 [ ] 6250

character code [ ] ASCII [ ] EBCDIC

[ ] printed copy (please, ONLY if it is impossible to send us machine-readable data)

II. CITATION INFORMATION

These data are [ ] published [ ] in press [ ] submitted [ ] in preparation [ ] no plans to publish

authors

title of paper

journal volume first-last pages year

Do you agree that these data can be made available in the database before they appear in print?

[ ] yes [ ] no, they should be made available only after publication (estimated date: _____)

Does the sequence which you are sending with this form include data that does **not** appear in the above citation?

[ ] no [ ] yes, from position _____ to _____ [ ] base pairs OR [ ] amino acid residues

(If your sequence contains 2 or more such spans, use the feature table in section IV to indicate their positions)

If so, how should these data be cited in the database?

[ ] published [ ] in press [ ] submitted [ ] in preparation [ ] no plans to publish

authors

address (if different from that given in section I)

title of paper

journal volume first-last pages year

List references to papers and/or database entries which report sequences overlapping with that submitted here.

first author	journal, vol., pages, year and/or database, accession number

C2.3/11.89

III. DESCRIPTION OF SEQUENCED SEGMENT

Wherever possible, please use standard nomenclature or conventions. If a question is not applicable to your sequence, answer by writing N.A.; if the information is relevant but not available, write a question mark (?).

What kind of molecule did you sequence? (check all boxes which apply)

[ ] genomic DNA [ ] genomic RNA [ ] virus [ ] provirus
[ ] cDNA to mRNA [ ] cDNA to genomic RNA
[ ] organelle DNA [ ] organelle RNA please specify organelle _____
[ ] tRNA [ ] rRNA [ ] snRNA [ ] scRNA

[ ] other nucleic acid (please specify) _____

[ ] peptide: [ ] sequence assembled by [ ] overlap of sequenced fragments [ ] homology with related sequence
[ ] other (please specify) _____
[ ] partial: [ ] N-terminal or [ ] C-terminal or [ ] internal fragment

length of sequence [ ] base pairs or [ ] amino acid residues

gene name(s) (e.g., *lacZ*)

gene product name(s) (e.g., beta-D-galactosidase)

Enzyme Commission number (e.g., EC 3.2.1.23)

gene product subunit structure (e.g., hemoglobin $\alpha_2\beta_2$)

The following items refer to the original source of the molecule you have sequenced.

organism (species) name (e.g., *Escherichia coli; Mus musculus*)

sub-species strain (e.g., K12; BALB/c)

name/number of individual or isolate (e.g., patient 123; influenza virus A/PR/8/34)

developmental stage [ ] germ line [ ] rearranged

haplotype tissue type cell type

The following items refer to the immediate experimental source of the submitted sequence.

name of cell line (e.g., Hela; 3T3-L1)

library (type; name) clone(s)

The following items refer to the position of the submitted sequence in the genome.

chromosome (or segment) name/number

map position units: [ ] genome % or [ ] nucleotide number or [ ] other _____

Using single words or short phrases, describe the properties of the sequence in terms of:

its associated phenotype(s);
the biological/enzymatic activity of its product;
the general functional classification of the gene and/or gene product
macromolecules to which the gene product can bind (e.g., DNA, calcium, other proteins);
subcellular localization of the gene product;
any other relevant information.

Example (for viral *erbB* nucleotide sequence): transforming capacity; EGF receptor-related; tyrosine kinase; oncogene; transmembrane protein.

IV. FEATURES OF THE SEQUENCE

Please list below the types and locations of all significant features experimentally identified within the sequence. **Be sure that your sequence is numbered beginning with "1."**

In the column marked	fill in
feature	type of feature (see information below)
from	number of first base/amino acid in the feature
to	number of last base/amino acid in the feature
bp	x, if your numbers refer to positions of base pairs in a nucleotide sequence
aa	x, if your numbers refer to positions of amino acid residues in a peptide sequence
id	method by which the feature was identified. **E** = experimentally; **S** = by similarity with known sequence or to an established consensus sequence; **P** = by similarity to some other pattern, such as an open reading frame
comp	x, if feature is located on the nucleic acid strand complementary to that reported here

Significant features include:

regulatory signals (e.g., promoters, attenuators, enhancers)
transcribed regions (e.g., mRNA, rRNA, tRNA). (indicate reading frame if start and stop codons are not present)
regions subject to post-transcriptional modificaton (e.g., introns, modified bases)
translated regions
extent of signal peptide, prepropeptide, propeptide, mature peptide
regions subject to post-translational modification (e.g., glycosylated or phosphorylated sites)
other domains/sites of interest (e.g., extracellular domain, DNA-binding domain, active site, inhibitory site)
sites involved in bonding (disulfide, thiolester, intrachain, interchain)
regions of protein secondary structure (e.g., alpha helix or beta sheet)
conflicts with sequence data reported by other authors
variations and polymorphisms

The first 2 lines of the table are filled in with examples.

If you think you will need more space than the table below provides, please photocopy this page before you fill it out.

Numbering for features on the sequence submitted here [ ] matches paper [ ] does not match paper						
feature	from	to	bp	aa	id	comp
EXAMPLE TATA box	1	8				
EXAMPLE exon 1	9	264				

C4.1/2.88

DDBJニュースレター申し込み書　　[ ] 新規　　[ ] 継続、訂正

必要事項を記入して下記の宛先までお送り下さい。
宛先：　411 三島市谷田1111、国立遺伝学研究所　遺伝情報センター
遺伝情報分析研究室　DDBJ係

ふりがな
氏名_____ 日付_____
ふりがな
所属_____ 電話_____
ふりがな
住所_____
（宛先を記したラベル2枚を同封下さい。）

DDBJニュースレター

[ ] 定期配布　　_____部　　[ ] 一時配布　　_____部

ＤＮＡ 及び 蛋白質配列解析ソフトウエア　マニュアル

1. UNIX用

- The Manual of the Flat Database and Sequence Analysis System for DNA and Proteins　[ ]
- The Manual of the Qanalys Sequence Analysis System for Molecular Evolution　[ ]

2. VAX/VMS用

- UWGCG及びIDEAS 利用の手引　[ ]
- Introduction to the Sequence Analysis Software Package of the University of Wisconsin Genetics Computer Group　[ ]
- User's Guide for the Protein Sequence Query Program of the Protein Identification Resource (PIR)　[ ]
- User's Guide for the Nucleic Acid Query Program of the Protein Identification Resource (PIR)　[ ]

データバンク運営に関するコメント

DNA，蛋白質データ配布申し込み書　[　] 新規　[　] 継続、訂正

必要事項を記入して下記の宛先までお送り下さい。[　] の中の。印はdefaultを意味します。　宛先：　411 三島市谷田1111、国立遺伝学研究所　遺伝情報センター
遺伝情報分析研究室　DDBJ係

ふりがな
氏名_____ 日付_____
ふりがな
所属_____ 電話_____
ふりがな
住所_____
（宛先を記したラベル2枚を同封下さい。）

DNA データ

[　] GenBank：　[。] MT　(6250 bpi, 2400ft; 1600 bpi, 2400ft × 3)
[。] 一時配布　[　] 定期配布（年　4 回）
[　] EMBL ：　(6250 bpi, 2400ft; 1600 bpi, 2400ft × 3)
[。] 一時配布　[　] 定期配布
[　] DDBJ ：　(6250 bpi, 600ft; 1600 bpi, 600ft)
[。] 一時配布　[　] 定期配布

蛋白質データ

[　] SwissProt: [。] 一般配布用 (6250 bpi, 1200 ft ; 1600 bpi, 2400ft)
[。] 一時配布　[　] 定期配布

（注）定期配布をお望みの方はあらかじめテープをお送り下さい。一時配布の場合は、あらかじめテープをお送り下さるか、もしくは使用後テープを送り返して下さい。

磁気テープ（9 Track ）フォーマット

Density:　[　] 1600 bpi　[。] 6250 bpi
使用できる最も高い Densityを指定してください。

Tape Label:　[。] unlabeled

Block size:　[　] 2400　[　] 3200　[　] 6400　[。] 12800 bytes　[　]

Record size:　[。] Fixed 80 bytes　[　] Variable

Character code:　[。] ASCII　（英小文字）　[　] EBCDIC

ソフトウエア配布申し込み書

必要事項を記入して下記の宛先までお送り下さい。
宛先： 411 三島市谷田1111、国立遺伝学研究所　遺伝情報センター
遺伝情報分析研究室　DDBJ係

ふりがな
氏名_____ 日付_____
ふりがな
所属_____ 電話_____
ふりがな
住所_____
（宛先を記したラベル2枚を同封下さい。）

[ ] DDBJ/EMBL/GenBank DNA Data Submission Form (MS-DOS 2HD or 2DD floppy)

[ ] FLAT Database and Sequence Analysis System for DNA and Proteins

UNIX(SUN OS/System V/BSD 4.2/BSD 4.3)で稼働するデータベースの検索、解析のための簡易プログラムパッケージです。検索はspeedよりflexibilityを考慮しました。DDBJ/GenBank, EMBL/SwissProt, PIR, PRFが処理可能です。解析プログラムは開発途上ですので十分とはいえません。使用するにはUNIXの知識が必要です。

[ ] Kermit and Tools　　5.25インチフロッピー 2HD 3枚をお送り下さい。
[ ] NEC PC9801　　[ ] _____

Kermitは ファイル転送用プログラムです。 NEC・PC9801版は 高エネルギー物理学研究所藤井氏により移植されたもので、端末エミュレーターとしては
- VT102
- TEKTRONIX 4014

をエミュレートします。9600 baudまで動作します。
日本語は Shift-JIS, JIS-83, JIS-78, UNIX コードが使用できます。ローマ字ー漢字変換フロントエンド ATOK も使用可能です。

Toolsはファイル転送用ツールです。

[ ] VT emulator　5.25インチフロッピー 2HD 3枚をお送り下さい。

東京大学医科学研究所伊藤氏作成したもので、DECUSソフトウエアライブラリーに登録されているPublic domain softwareです。NEC PC-9801/XAで以下の端末をエミュレートする。
- VT52/VT80E/VT100/VT220/VT282
- TEK4010/TEK4012/{TEK4014}/{VT55}/VT125/VT240/VT284モノクログラフィック
- {TEK4027A}/{GIGI}/VT241/VT246　カラーグラフィック

VTシリーズ端末エミュレーターとしては完璧である。日本語変換としてはNEC標準の文節変換が使用できます。（NECDIC.DRV, NECDIC.SYSを使用します。）
VAX/VMSを使用する際だけでなくグラフィック端末エミュレーターにもなりますのでUNIX用としても有用です。ファイル転送用Kermitは、MS-DOS Generic版もPC98用の実行プログラムが付属していますが、先のPC98版の使用をお薦めします。
マニュアルは、印刷物としては配布いたしません。ファイルを出力して下さい。（印刷したものを入手したい方は、伊藤氏まで問い合わせ下さい。）

国立遺伝学研究所DNAデータベースオンライン利用申請書

平成　　年　　月　　日

国立遺伝学研究所長　　殿

貴研究所のＤＮＡデータベースオンライン利用について下記のとおり申請します。なお、それらの利用にあたっては、「国立遺伝学研究所ＤＮＡデータベース等利用規則」を遵守します。

記

①申請区分	□新規 □継続	②利用期間	平成　年　月　日～平成　年　月　日	
※アカント種別、その他				
※ＵＩＤ		※ＧＩＤ		
※パスワード		※ユーザネーム		
③利用申請者	職　名		(ﾛｰﾏ字) 氏　名	㊞
	(英訳) 所　属			
	(英訳) 所在地	〒□□□－□□　　☎　　(　　)		
④利用目的				
⑤支払責任者	職　名		氏　名	㊞
	所　属			
	所在地	〒□□□－□□　　☎　　(　　)		
⑥経理責任者	職　名		氏　名	㊞
	所　属			
	所在地	〒□□□－□□　　☎　　(　　)		
⑦利用見込額	円	⑧支出科目	□国立学校校費 □附属病院費 □文部省科学研究費 □研究所校費 □国立学校受託研究費 □その他(　　)	
※備　考				

注意　１．※印欄は記入しないで下さい。　２．裏面の記入要領をよく読んで下さい。

記　入　要　領

① 申請区分　該当する事項にチェックして下さい。なお、「継続」とは、利用期間終了後、引き続き利用申請する場合をいいます。

② 利用期間　利用期間は、一会計年度内ですので、その間の利用期間を記入して下さい。

③ 利用申請者　職　名　教授、助教授、講師、助手、研究員等と記入して下さい。なお、大学院生は、「博士」、「修士」の別を記入して下さい。

所　属　申請者が所属する大学、学部、学科又は研究所等の名称を記入して下さい。

なお、大学院生は、研究科名、専攻名まで記入して下さい。

氏　名　上段に氏名をローマ字で名、姓の順に記入して下さい。

所在地　所属の住所を記入して下さい。なお、所属がない場合には、現住所を記入して下さい。

④ 利用目的　当研究所ＤＮＡデータベース等を利用を必要とする研究テーマを記入して下さい。

⑤ 支払責任者　1)　申請者が支払うべき利用負担金については、その支払いに責任をもてる者を記入して下さい。

2)　支出科目が科学研究費の場合は、研究費の配分を受けている者を記入して下さい。

3)　所属及び所在地が申請者と同じときは、「利用申請者に同じ」と記入して下さい。

⑥ 経理責任者　1)　予算執行の法的責任を有する事務担当者を記入して下さい。

例えば、事務〔部〕長、会計〔経理〕課長、会計〔経理〕係長等

2)　所属及び所在地が申請者又は支払責任者と同じときは、「利用申請者に同じ」又は「支払い責任者に同じ」と記入して下さい。

⑦ 利用見込額　利用料金の見込額を記入して下さい。

⑧ 支出科目　該当する事項にチェックして下さい。ただし、「その他」の場合は私費等その経費の名称を記入して下さい。

昭和　　年　　月　　日

国立遺伝学研究所DNAデータベース等利用　終了／中止／承認内容変更　届

国立遺伝学研究所長 殿

ユーザネーム								
職　名		氏名	㊞					
所　属								

下記により、ＤＮＡデータベース等利用　を終了／を中止／の承認内容を変更　したのでお届けします。

記

終了／中止／変更　理　由			
終了／中止／変更　年月日	昭和　　年　　月　　日	備考	

Mailing Addresses for Inquiries

If you have any inquiry, please send mails to the following addresses (...@ddbj.nig.ac.jp).

postmaster	about mails, including address representation.
ddbj	about DDBJ activities
ddbjsub	data submission to DDBJ
genbank	inquiries to GenBank
gbsub	data submission to GenBank
embl	inquiries to EMBL
emblsub	data submission to EMBL

Mails to genbank, and embl will be forwarded to them.

DNA Data Bank of Japan 所在地

411 三島市谷田1111
国立遺伝学研究所
遺伝情報研究センター
遺伝情報分析研究室内 DDBJ
電話 0559-75-0771 (内) 647
FAX: 0559-75-6040
E-mail: ddbj@ddbj.nig.ac.jp

DNA Data Bank of Japan スタッフ

		(E-mail:)
Manager/Database administrator	遺伝情報分析研究室	宮沢三造 (smiyazaw)
Scientific Reviewer	遺伝情報分析研究室	林田秀宜 (hhayashi)
関連事務担当	DDBJ 非常勤	堀江元乃 (ddbj-secretaries)
		鈴木成子 (ddbj-secretaries)
		(@ddbj.nig.ac.jp)