ＤＤＢＪニュースレター　　No. 8　　1989年 3月

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

目　　次

付属資料

計算機接続回線：DDBJnewsでloginして各種情報を得ることができます。

電話回線

外線　M-380Q/UTS (Unix System V Release 2)

0559-75-6036　　CCITT 2400/1200 bpi, MNP error correction

6037　　CCITT 2400/1200 bpi, MNP error correction

MicroVAX II/VMS

0559-75-6038　　CCITT 2400/1200 bpi, MNP error correction

内線　FACOM 380Q/UTS

0559-75-0771 : 676　　CCITT 2400/1200 bpi

677　　CCITT 2400/1200 bpi

678　　CCITT 2400/1200 bpi

679　　CCITT 2400/1200 bpi

DDX-P address : 522-5127　（網間接続の場合は 163-060-522-5127）

- 回線初期設定　：　Full duplex, Remote echo, No party, 8 bit code
  1 start bit, 1 stop bit,　Xon/Xoff
- UTS の場合は、Break 信号により2400 bpi→1200→300 →2400を切り換えます。
- VMS の場合は、autoband 設定により自動切り換えですので、＜CR＞を数回送ってください。

UTS login 時の注意

- usernameは必ず英小文字
  (Initial of first name + first 7 characters of last name)
- terminal type はPC9801のmsdos がdefault です。vt100 その他多くの端末がサポートされています。

VAX login 時の注意

- terminalは DEC端末(VT100,VT200,...)か、dumb端末のみサポートします。

UTS, VMS　その他の計算機(SUN,IRIS) は互いにremote login可能です。
UTS で "gentinfo" コマンドを利用し、その他必要な情報を得て下さい。

ＤＤＢＪ活動報告と今後の計画

遺伝情報分析研究室
ＤＤＢＪ

以下簡単に1988年度のDDBJの活動を報告する。詳しくは、各章を参照願いたい。

1) DNAデータ入力
1988年 7月に 3版（230エントリー、345,850塩基）、1989年 1月に 4版（302エントリー、535,985塩基）をリリースした。 Accession numberが D で始まるものは DDBJ が入力したものです。
2) DNAデータベースのための国際諮問委員会（1988年 2月と1989年 2月）参加
3) DDBJ, EMBL, GenBank共通の New feature table のデザインに参加
DDBJ/EMBL/GenBank Feature Table: Definitionを 1988年 9月に完成させた。
New feature table の使用は 1989年後半になる予定。
4) DDBJ, EMBL, GenBankの実務協議（1988年 9月）に参加
データバンクはデータの増加と質の変化（ゲノム解析）に対応するためデータベースの再構築（関係データベースへの移行）を行いつつある。本会議の主な議題は関係データベース設計の詳細、データ入力支援ソフトウエアの仕様、CD-ROMによるデータ配布、データの質的向上（curatorシステムの採用）等についてであった。DDBJにおける関係データベースへの移行は 1989年後半になる予定。
5) DDBJ利用初進者講習会開催（1988年 6月）
DDBJの利用（データ提出の方法及び検索システムの使用法）に関し初進者講習会を開催した。20名の方に参加いただいた。
6) DNA、蛋白質データベースの配布
1988年の配布総数は、磁気テープ 612本　フロッピー 572枚である。
7) マニュアルの作成、配布
DDBJで利用可能なソフトウエアのマニュアルの作成、配布した。
8) ソフトウエアの配布
従来から配布している端末エミュレーターに加え、 UNIXシステムの上で稼働する FLAT データベース検索解析プログラムパッケージの配布を開始した。配布するものはDDBJ計算機の上で稼働しているものとほぼ同じである。
9) 学会デモンストレーション
DDBJの利用（データ提出の方法及び検索システムの使用法）に関し、日本癌学会（1988年 9月）、日本生化学会（1988年10月）、日本分子生物学会（1988年12月）でデモンストレーションを行った。

なお、NIHのGenBank担当官 Dr. J. Cassatt氏と EMBLの Dr. C. Sanderが 1988年10月来日の折り DDBJを訪問し、Dr. Cassattには "Nucleic Acid Databases - Vision for the Future"、Dr. Sanderには "EMBNet: Network for Molecular Biology in Europe"という題で研究所で話していただいた。Dr. Sander氏の講演についてはその要旨を掲載したのでご覧いただきたい。

なお、DDBJの 1989 年度の主な計画は

1) GenBank が収集を担当しているデータで日本で生産されるものは DDBJ で入力
2) 関係データベースへの移行：GenBank, EMBL と同じ関係データベースを採用
2) New feature table への移行：DDBJ/EMBL/GenBank 共通
4) ネットワークの構築
4-1) 米国の ESNET, Internet と TCP/IP で接続
4-2) 学術情報ネットワークへの接続：関係機関と TCP/IP もしくは DECnet でネットワーク構築

である。スタッフが十分でなく、データ入力、管理で手いっぱいである。御理解頂きたい。

DNAデータ入力：DDBJ Release 3 及び Release 4

遺伝情報分析研究室
宮沢三造、林田秀宜

1986年12月に DNAデータを収集し始めて以来、半年毎にリリースしてきた。1988年 7 月に 3版、1989年 1 月に 4版をリリースした。 4版は、302 エントリー、535,985 塩基を含んでいる。過去リリースした各版のエントリー数及び塩基数を Table. 1 に示す。 データバンクは統一した Accession number を使用しているので GenBank, EMBL データベースでも Accession numberが D で始まるものは DDBJ が入力したものです。

1987年 7月以来1988年までの 1年間で約 250 kb 収集したことになる。同期間に 3データバンク (DDBJ, EMBL, GenBank) で 8 Mb 収集した。言い替えるとDDBJはこの一年間で全世界で収集した塩基数の約 1/30 を収集したことになる。この数は一見あまりに少ないようであるが、DDBJにおけるスタッフ数を GenBank 及び EMBL における人数と比較してみると当然とも思える数である。現在DDBJでは、2 fuculty position, 注釈者として 0.5 full time employees (FTEs), Reviewerとして 0.2 FTEs、秘書一名、及び入力業務一名で
1) データ収集、管理　2) データ配布　3) データベースのオンラインによる利用のサポート　4) 解析ソフトウエアの開発、整備　5) ニュースレターの発行　6) 講習会の開催
を行っている。ちなみに GenBank では30 FTEsで 1) の仕事を、 EMBL では 20 FTEs のスタッフで 1) と 2) を行っている。DDBJが入力数を増やすには、研究者の皆様の協力を得なくてはならない。 DNA データを発表した際には必ずデータを Submission して欲しい。" DNA 配列データ提出のお願いと DDBJ, EMBL, GenBank へのデータ提出について" の章を参照して欲しい。

現在DDBJは、日本で出版される雑誌を主にデータ収集している。 Table. 2 は収集対象雑誌と各雑誌で見いだされた DNA 配列を含む論文数を示している。そして予測されるように2.3の雑誌に集中しているとは言っても、数は少なく J. Biochem. Tokyo で約 20 - 25 /year 程度である。全てを集めても約 30 -40/year 程度である。BIOSIS データベースによると1987年度で1279論文発表され、その内 148編は日本の研究機関からの発表とのことである。DDBJは過去一年で70論文処理した。つまり日本の研究機関から発表される約 1/2、全世界で発表される論文の内 1/20 を処理したことになる。BIOSIS は全ての論文をカバーしているとは言えないだろう。よって先の 1/30 という数字が妥当かもしれない。

現在のデータ入力システムは、Fig. 1 に示されている。また、各データベースのエントリー数と塩基数の変遷を Fig. 2 に示した。

なお、DDBJ担当の各学術雑誌には投稿者へ Data Submission Form を送付してくれるよう依頼の手紙を出した。詳しくは "DNA データ収集に関する学術雑誌との協力関係について" を参照願いたい。

Table 1. The numbers of entries and bases in each release of the DDBJ database

Release	Date	Entries	Bases
1	07/87	66	108,970
2	01/88	142	199,392
3	07/88	230	345,850
4	01/89	302	535,985

Table 2. Journals scanned by the DDBJ and the number of papers found to include original DNA sequences; this data was collected in January 30, 1989.

Journals published in Japan:		entries	papers	
Agricul Biol Chem	Vol. 44(01)80-46(12) 1982	0	0	
	Vol. 47(01) - 47(12) 1983	1	1	
	Vol. 48(01) - 48(12) 1984	0	0	
	Vol. 49(01) - 49(12) 1985	2	2	
	Vol. 50(01) - 50(12) 1986	3	3	
	Vol. 51(01) - 51(12) 1987	12	11	
	Vol. 52(01) - 52(10) 1988	14	12	
Cell Struc Funct	Vol. 11(01) - 11(04) 1986	0	0	
	Vol. 12(01) - 12(04) 1987	0	0	
	Vol. 13(01) - 13(05) 1988	0	0	
Chem Pharm Bull	Vol. 34(12) - 34(12) 1986	0	0	
	Vol. 35(01) - 35(12) 1987	0	0	
	Vol. 36(01) - 36(10) 1988	0	0	
Devel Growth Diff	Vol. 28(01) - 28(06) 1986	0	0	
	Vol. 29(01) - 29(06) 1987	0	0	
	Vol. 30(01) - 30(06) 1988	0	0	
J Biochem Tokyo	Vol. 99(01) - 99(06) 1986	11	8	
	Vol.100(01) -100(06) 1986	27	14	
	Vol.101(01) -101(06) 1987	15	6	
	Vol.102(01) -102(06) 1987	28	14	
	Vol.103(01) -103(06) 1988	50	15	
	Vol.104(01) -104(06) 1988	13	7	
Jpn J Cancer Res	Vol. 77(01) - 77(12) 1986	0	0	
	Vol. 78(01) - 78(12) 1987	1	1	
	Vol. 79(01) - 79(10) 1988	1	1	
Jpn J Genet	Vol. 50(01)75-60(06) 1985	0	0	
	Vol. 61(01) - 61(06) 1986	10	2	
	Vol. 62(01) - 62(06) 1987	5	5	
	Vol. 63(01) - 63(05) 1988	1	1	
Microbiol Immunol	Vol. 31(02) - 31(12) 1987	3	2	
	Vol. 32(01) - 32(10) 1988	1	1	
Plant Cell Physiol	Vol. 28(01) - 28(08) 1987	2	2	
Zool Sci	Vol. 3(01) - 3(06) 1986	0	0	
	Vol. 4(01) - 4(06) 1987	0	0	
	Vol. 5(01) - 5(04) 1988	0	0	
Nippon Ika Daigaku Zasshi	Vol. 54 1987	2	2	Not scanned
Journals published outside of Japan:				
J Gen Virol	Vol. 68(03) - 68(12) 1987	38	27	
	Vol. 69(01) - 69(11) 1988	53	33	

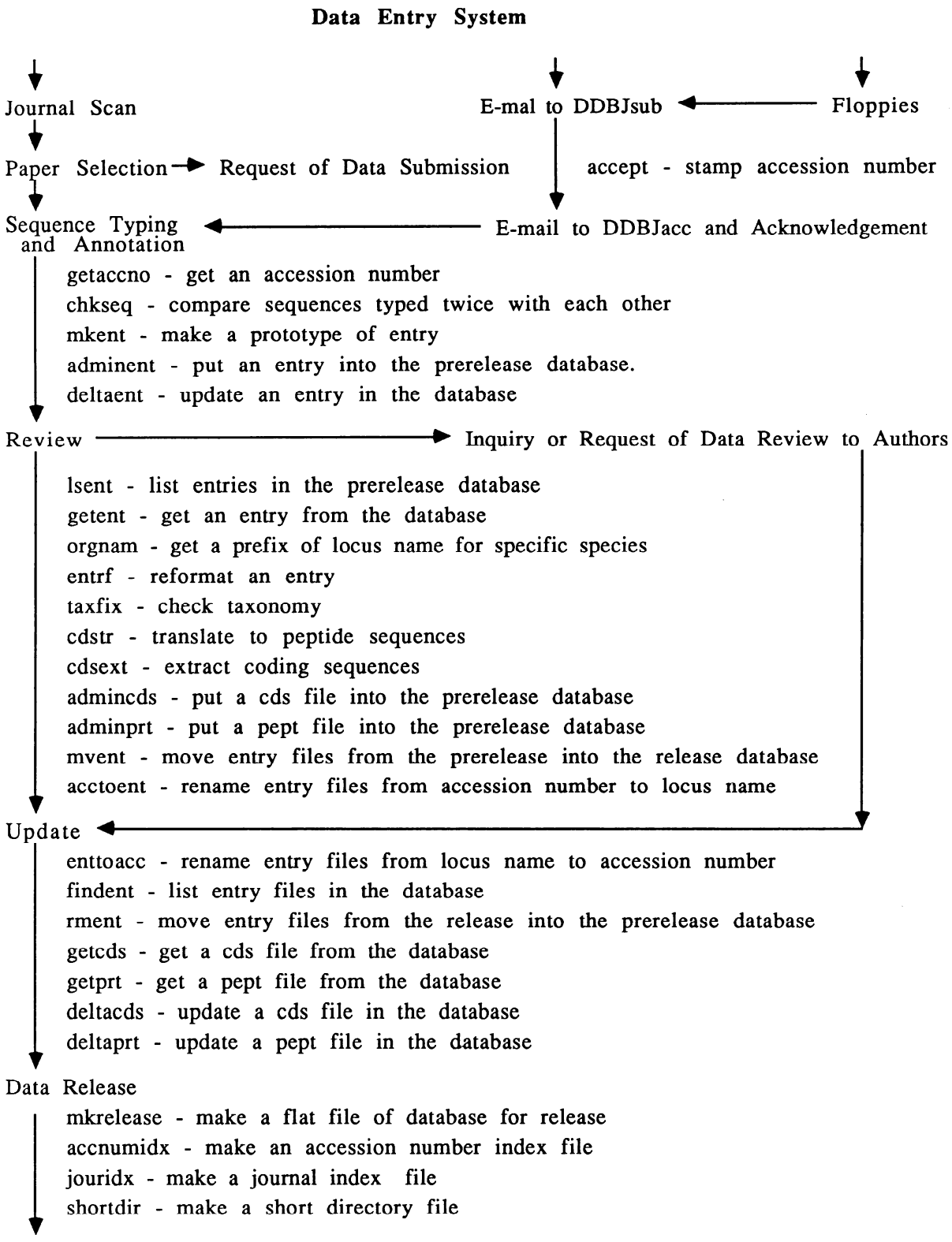
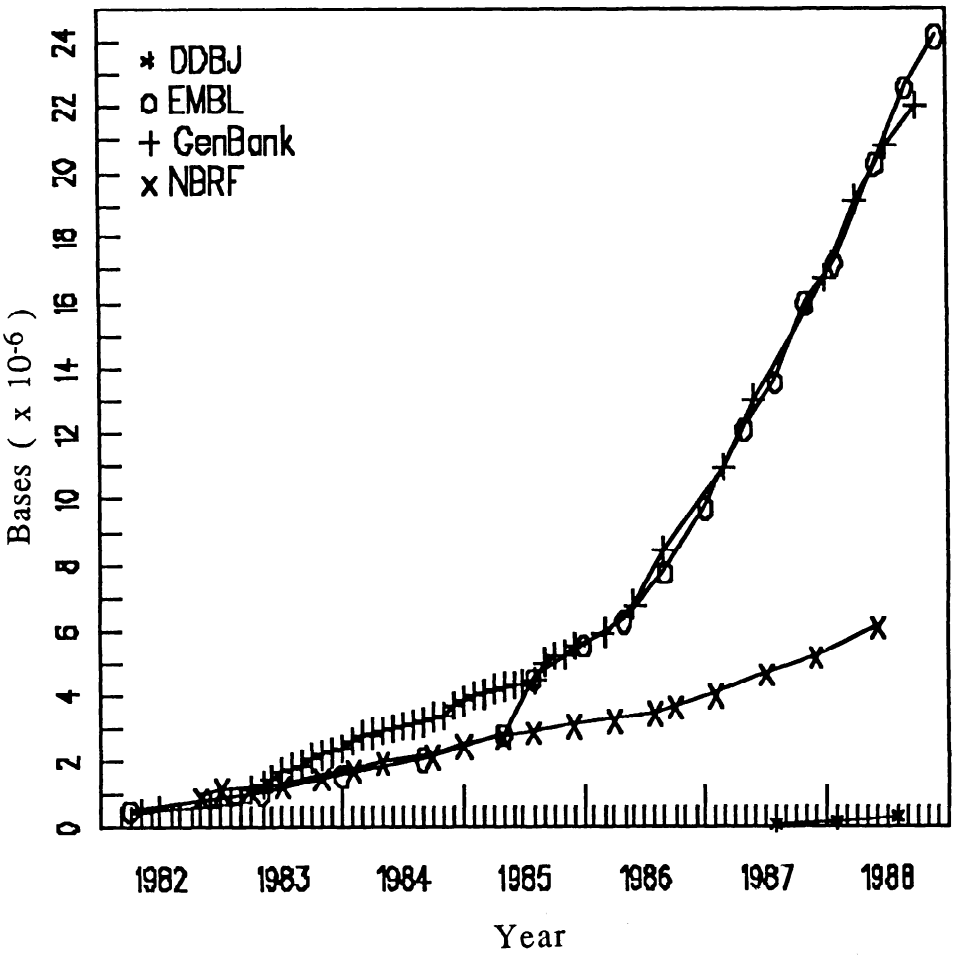
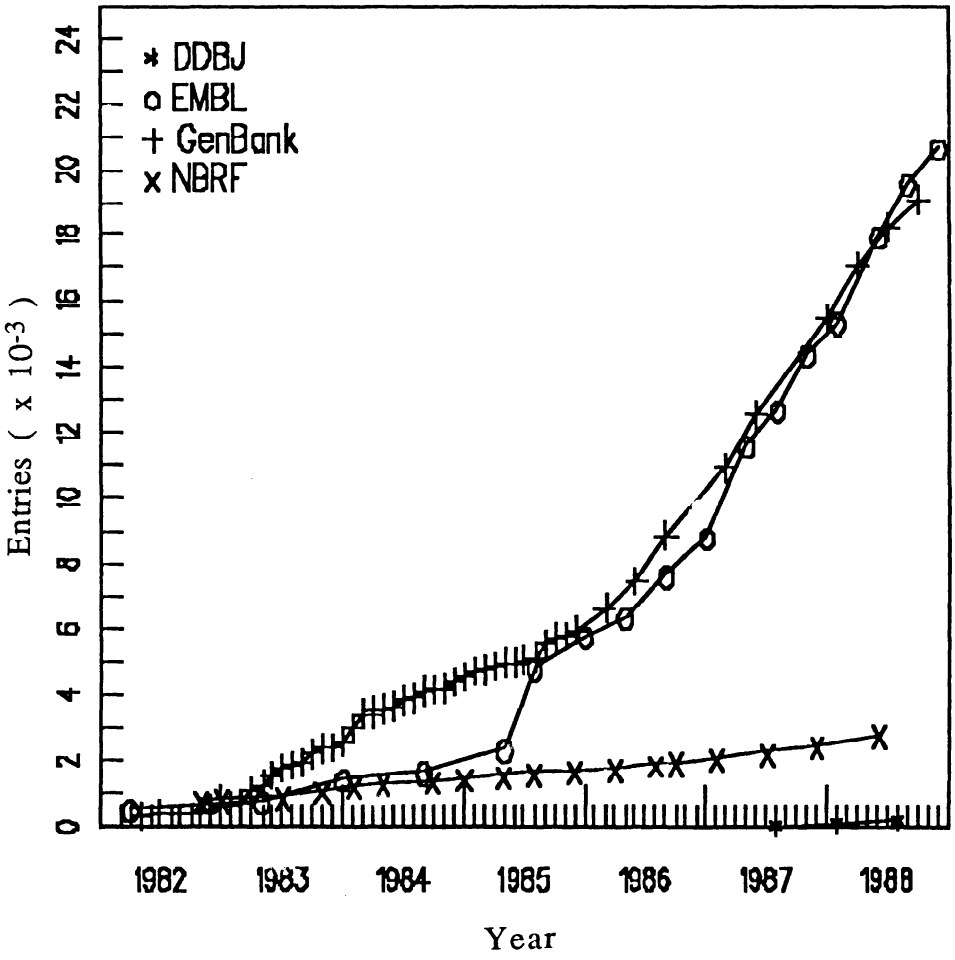
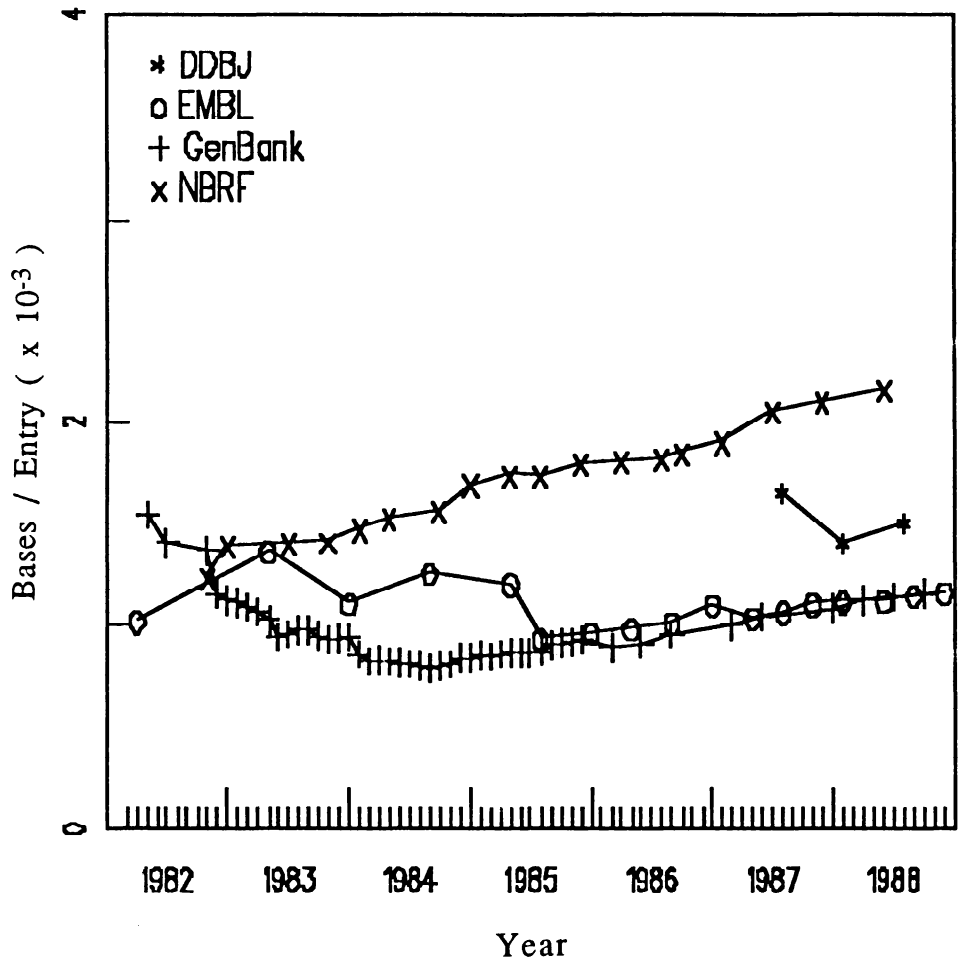
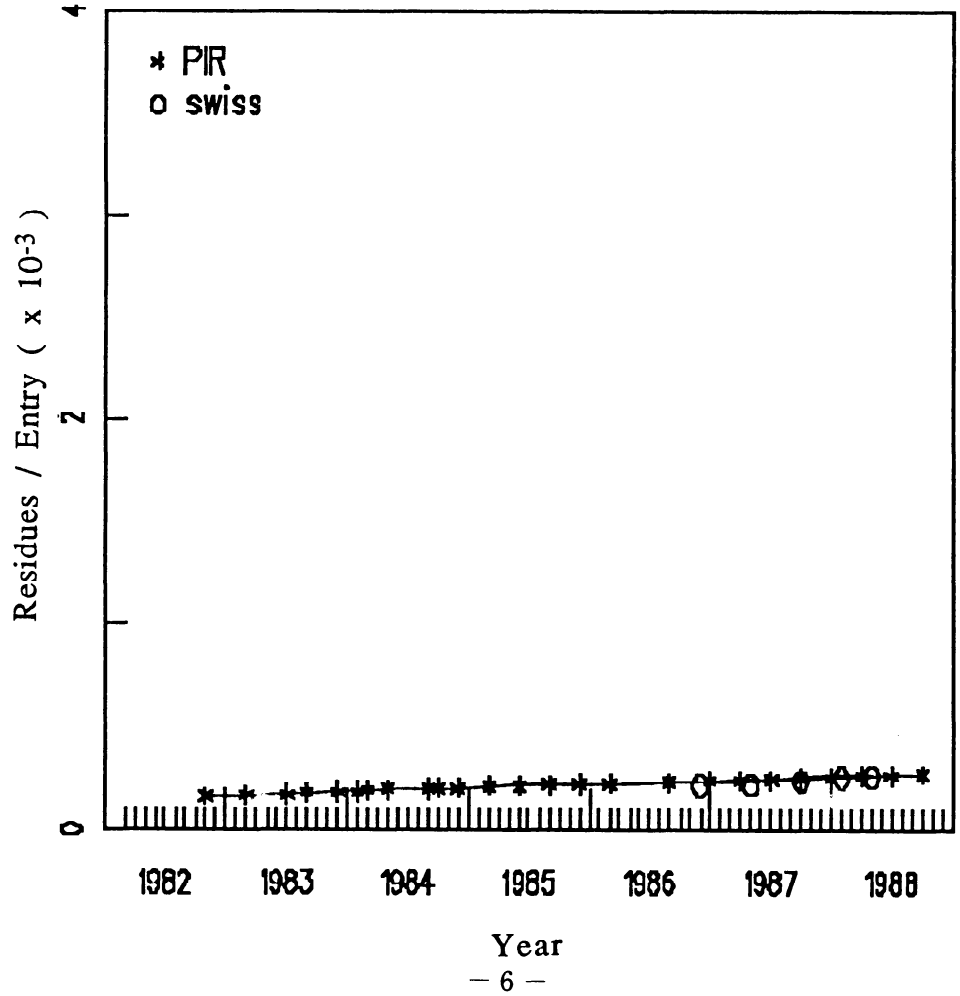


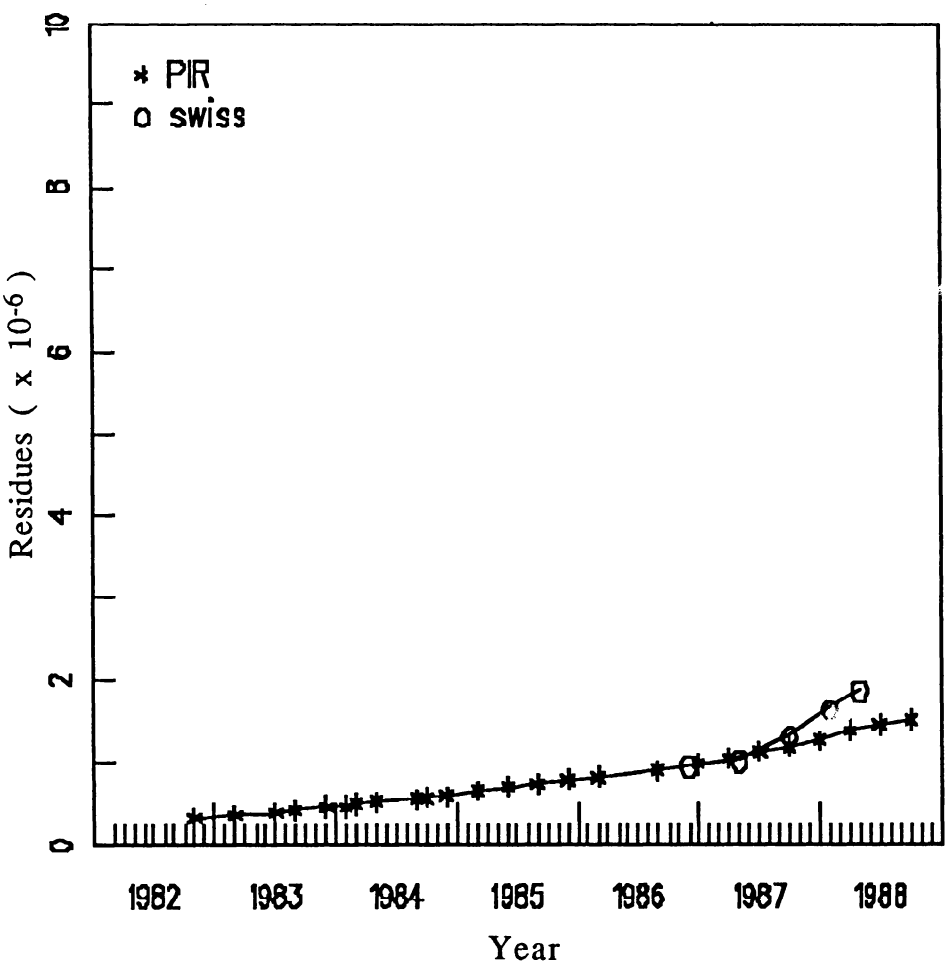
FIGURE 1 Data flow and commands that are used at each step of data entry

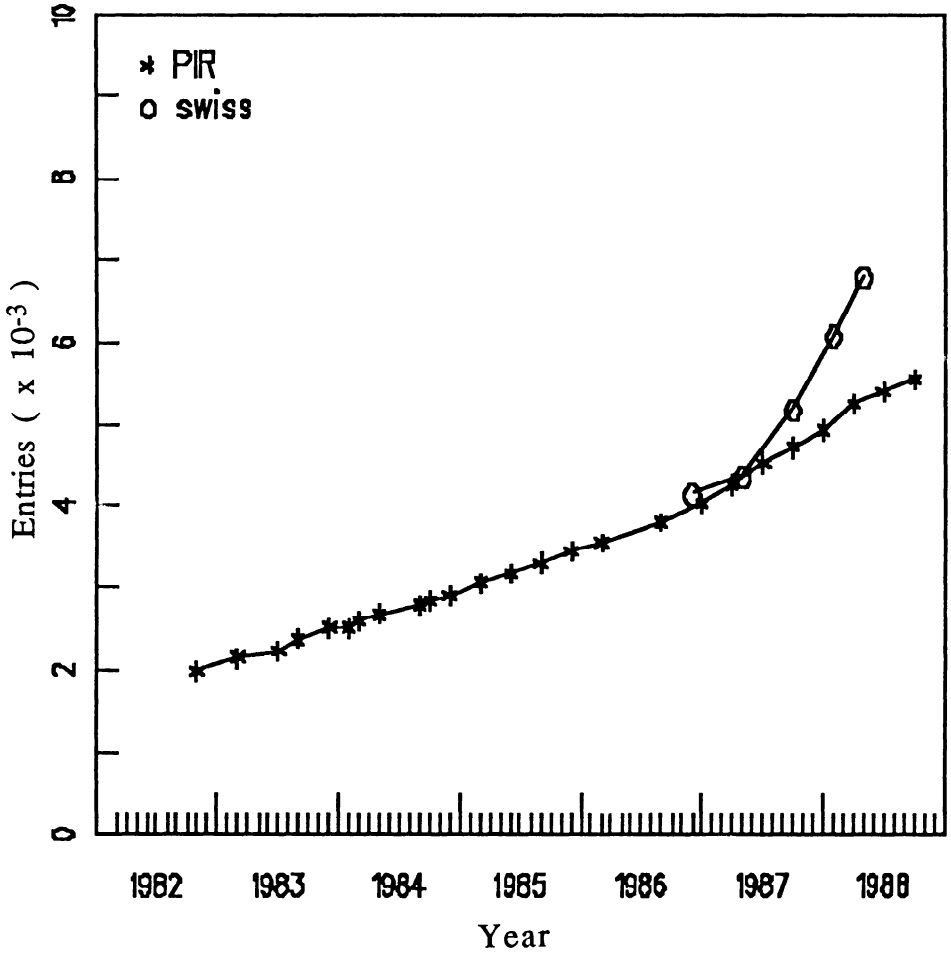

* DDBJ
o EMBL
+ GenBank
x NBRF
Bases (x 10-6)
0 2 4 6 8 10 12 14 16 18 20 22 24
1982 1983 1984 1985 1986 1987 1988
Year


* DDBJ
o EMBL
+ GenBank
x NBRF
Entries (x 10-3)
0 2 4 6 8 10 12 14 16 18 20 22 24
1982 1983 1984 1985 1986 1987 1988
Year


* DDBJ
o EMBL
+ GenBank
x NBRF
Bases / Entry (x 10-3)
0
2
4
1982
1983
1984
1985
1986
1987
1988
Year


* PIR
o swiss
Residues / Entry (x 10-3)
0
2
4
1982
1983
1984
1985
1986
1987
1988
Year


* PIR
o swiss
Residues (x 10-6)
0
2
4
6
8
10
1982
1983
1984
1985
1986
1987
1988
Year


* PIR
o swiss
Entries (x 10-3)
0
2
4
6
8
10
1982
1983
1984
1985
1986
1987
1988
Year

DNA データフォーマットの変更について

宮澤　三造

過去一年間配列データの書式が以下のように若干変更された。

GenBank Release 54 (12/87)、DDBJ Release 4 (1/89)

- Features tableを改変しSites tableはNew features table に吸収された。

GenBank Release 55 (3/88)、 DDBJ Release 4 (1/89)

- LOCUSレコードの書式が変わった。生物グループ名（3 文字表現）を含むこと、また日付の書式がEMBL方式に変わった。
- データ注釈のレベルを表示するSTANDARDレコードがREFERENCEレコードの一部として追加された。

また、今後以下のような変更が計画されている。

- 論文雑誌名の省略方法を変更する予定。
  データバンク (DDBJ, GenBank, EMBL, PIR)は National Library of Medicine (USA)と共同で省略名の標準化を進めている。
- EMBL は配列データで U の代わりに T を使用する予定である。
- New features table の使用　(DDBJ/EMBL/GenBank 共通)

最も重要な変更はデータ注釈に関するNew features table の使用である。従来のFeatures table は GenBank方式、EMBL方式ともに欠陥があった。重要な情報の多くが Keywordsを用い的確に表現することが不可能であるため、コメントとして書き表す以外に手がなく結果としてプログラムでは解読が困難となった。これは、現在用いているFeatures tableには科学の急速な進展にともなって明らかになった知見を取り込むだけの拡張性に欠けている側面があったためと思われる。一方、現在のFeatures tableでは EMBL formatと GenBank formatの書式変換が困難なものとなる。このような反省の上に立って New features tableが DDBJ, EMBL, GenBank 共同で作成された。非常に困難な作業であったため、数年に渡る議論を必要とし、1988年 9月に完成した。 New features table definition manual を希望者に配布するので DDBJまで申し込み書を送付願いたい。以下簡単にその特徴を示す。

Old and New Feature Tables: GenBank and EMBL entries

GenBank Old Entry

```
LOCUS       HUMPALB       614 bp ss-mRNA            PRI       15-JUN-1988
DEFINITION  Human prealbumin mRNA, complete cds.
ACCESSION   M10605
KEYWORDS    prealbumin.
SOURCE      Human liver, cDNA to mRNA, clone PA7.
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Vertebrata; Tetrapoda; Mammalia;
            Eutheria; Primates; Anthropoidea; Hominoidea; Hominidae.
REFERENCE   1  (bases 1 to 614)
  AUTHORS   Wallace,M.R., Naylor,S.L., Kluve-Beckerman,B., Long,G.L.,
            McDonald,L., Shows,T.B. and Benson,M.D.
  TITLE     Localization of the human prealbumin gene to chromosome 18
  JOURNAL   Biochem Biophys Res Commun 129, 753-758 (1985)
  STANDARD  simple staff_review
COMMENT     Draft entry and sequence in computer readable form for [1] kindly
            provided by M.R.Wallace, 26-DEC-1985.
FEATURES        from  to/span     description
    pept          26      469     prealbumin
    sigp          26       85     prealbumin signal peptide
    matp          86      466     prealbumin
    mRNA    <      1      614     prealb mRNA
    refnumbr       1        1     numbered 1 in [1]; zero not used
BASE COUNT      148 a    162 c    155 g    149 t
ORIGIN      247 bp upstream of AluI site; chromosome 18.
```

GenBank New Feature Table

```
FEATURES             LOCATION/QUALIFIERS
    sig_peptide      26..85
    mat_peptide      86..466
                     /product="prealbumin"
    CDS              26..469
                     /product="prealbumin"
    mRNA             <1..614
BASE COUNT       148 a    162 c    155 g    149 t
ORIGIN      247 bp upstream of AluI site; chromosome 18.

----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+
        10        20        30        40        50        60        70
```

EMBL Old Entry

```
ID   ACAC01     standard; DNA; 1571 BP.
XX
AC   V00002;
XX
DT   26-JAN-1984  (species spelling)
DT   15-JUN-1983  (first entry)
XX
DE   Acanthamoeba castelani gene encoding actin I.
XX
KW   actin.
XX
OS   Acanthamoeba castellani (amoeba, amibe, Amoebe)
OC   Eukaryota; Protozoa; Rhizopoda.
XX
RN   [1]    (bases 1-1571; enum. 1 to 1571)
RA   Nellen W., Gallwitz D.;
RT   "Actin genes and actin messenger RNA in Acanthamoeba
RT   castellani. Nucleotide sequence of the split actin
RT   gene I";
RL   J. Mol. Biol. 159:1-18(1982).
XX
FH   Key         From     To       Description
FH
FT   TRANSCR       90  >1443       primary transcript
FT   TRANSCR       91  >1443       primary transcript (alternate start)
FT   MSG           90    451       exon 1
FT   MSG           91    451       exon 1 (alternate start)
FT   CDS          137    451       actin part 1
FT   IVS          452    580       intron I
FT   MSG          581  >1443       exon 2
FT   CDS          581   1390       actin part 2
XX
SQ   Sequence  1571 BP;  313 A;  535 C;  389 G;  334 T;
```

EMBL New Feature Table

```
XX
FH   Key             Location/Qualifiers
FH
FT   prim_transcript one-of(90,91)..>1443
FT   exon            one-of(90,91)..451
                     /number=1
FT   intron          452..580
                     /number=1
FT   exon            581..>1443
                     /number=2
FT   CDS             join((137..451),(581..1390))
                     /product="actin I"
XX
SQ   Sequence  1571 BP;  313 A;  535 C;  389 G;  334 T;
----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+
       10        20        30        40        50        60        70
```

New Feature Table: Backus-Naur form

Feature table is an optional part of an entry. Full entry syntax is specified elsewhere.

feature_table ::= <feature_table_header><feature_table_body>

feature_table_header ::= FH Key Location/Qualifiers |
FEATURES Location/Qualifiers

feature_table_body ::= <feature> | <feature_table_body><feature>

At least one feature is required.

feature ::= <feature_key><feature_details>

Key is required, location required, qualifier list optional

feature_key ::= <symbol> | –

feature_details ::= <location><qualifier_list> | <location>

There exists a table of legal keys. "-" is a placeholder for no key.

location ::= <absolute_location> | <feature_name> | <literal_sequence> |
<functional_operator>(<location_list>)

absolute_location ::= <local_location> | <path> : <local_location>

path ::= <database> :: <primary_accession> | <primary_accession>

feature_name ::= <path>:<feature_label> | <feature_label>

feature_label :== <symbol>

local_location ::= <base_position> | <between_position> | <base_range>

location_list ::= <location> | <location_list>,<location>

functional_operator ::= <symbol>

base_position ::= <integer> | <low_base_bound> | <high_base_bound> |
<two_base_bound>

low_base_bound ::= > <integer>

high_base_bound ::= < <integer>

two_base_bound ::= <base_position>.<base_position>

between_position ::= <base_position>^<base_position>

base_range ::= <base_position>..<base_position>

database ::= <symbol>

primary_accession ::= <symbol>

literal_sequence ::= <sequence_character> | <literal_sequence><sequence_character>

sequence_character ::= a | b | c | d | g | h | k | m | n | r | s | t | u | v | w | y

qualifier_list ::= <qualifier> | <qualifier_list><qualifier>

qualifier ::= /<qualifier_name> | /<qualifier_name>=<value>

qualifier_name ::= <symbol>

value ::= <simple_value> | (<value_list>) | (<tagged_value_list>)

simple_value ::= <integer> | <location> | <reference_number> | "<text_string>" |
<symbol>

value_list ::= <value> | <value_list>,<value>

tagged_value_list ::= <tagged_value> | <tagged_value_list>,<tagged_value>

tagged_value ::= <tag>:<value>

tag ::= <symbol>

reference_number ::= [<unsigned_integer>]

symbol ::= <letter> | <symbol><symbol_character> | <symbol_character><symbol>

text_string ::= <string_character>| <text_string><string_character>

unsigned_integer ::= <digit> | <unsigned_integer><digit>

integer ::= <unsigned_integer> | - <unsigned_integer>

string_character ::= <letter> | <digit> | <punctuation> | ""

symbol_character ::= <up_case_letter> | <low_case_letter> |<digit> | _ | - | ' | *

letter ::= <up_case_letter> | <low_case_letter>

up_case_letter ::= A | B| ... | Z

low_case_letter ::= a | b | ... | z

digit ::= 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9

punctuation ::= <space> | ! | # | $ | % | & | ' | (|) | * | + | , | - | . | / | : | ; | < | = | > | ? | @ | [
| \ |] | ^ | _ | ` | { | <bar> | } | ~

bar ::= |

space ::= ascii 32

Feature key relationship tree

A. misc_feature

1. misc_difference
   a) conflict
   b) unsure
   c) old_sequence
   d) mutation
   e) variation
   f) allele
   g) modified_base

2. misc_signal
   a) promoter
      1) CAAT_signal
      2) TATA_signal
      3) -35_signal
      4) -10_signal
      5) GC_signal
   b) RBS
   c) polyA_signal
   d) enhancer
   e) attenuator
   f) terminator
   g) rep_origin

3. misc_RNA
   a) prim_transcript
      1) precursor_RNA
         a)mRNA
         b) 5'clip
         c) 3'clip
         d) 5'UTR
         e) 3'UTR
         f) exon
         g) CDS
            1) sig_peptide
            2) transit_peptide
            3) mat_peptide
         h) intron
         i) polyA_site
         j) rRNA
         k) tRNA
         l) scRNA
         m) snRNA

4. repeat_region
   a) repeat_unit
   b) LTR
   c) satellite

5. misc_binding
   1) primer_bind
   2) protein_bind

6. misc_recomb
   a) cellular
   b) iDNA
   c) insertion_seq
   d) transposon
   e) provirus
   f) virion

7. misc_structure
   a) stem_loop
   b) D-loop

Qualifiers for Feature Keys

Qualifiers	Examples
/anticodon=(pos: ,aa:)	/anticodon=(pos:34..36, aa:Phe)
/bound_moiety=	/bound_moiety="repressor"
/citation=	/citation=[3]
/codon=(seq: ,aa:)	/codon=(seq:"ttt", aa:Leu)
/codon_start=	/codon_start=213
/cons_splice=	/cons_splice=(5'site:YES, 3'site:NO)
/direction=	/direction=LEFT
/EC_number=	/EC_number=1.1.2.4
/evidence=	/evidence=EXPERIMENTAL
/frequency=	/frequency=0.85
/function=	/function="essential for recognition of cofactor"
/gene=	/gene="ilvE"
/label=	/label=Alb1_exon1
/mod_base=	/mod_base=m5c
/note=	/note="This qualifier is equivalent to a comment."
/number=	/number=4
/organism=	/organism="Homo sapiens"
/partial	/partial
/phenotype=	/phenotype="erythromycin resistance"
/product=	/product="catalase"
/pseudo	/pseudo
/rpt_family=	/rpt_family="Alu"
/rpt_type=	/rpt_type=INVERTED
/rpt_unit=	/rpt_unit=Alu_rpt1
/standard_name=	/standard_name="dotted"
/transl_except=(pos: ,aa:)	/transl_except=(pos:213..216, aa:Trp)
/type=	/type="W64msw"
/usedin=	/usedin=X10087:proteinx

DNA データ収集に関する学術雑誌との協力関係について

以下の学術雑誌の編集幹事あて、論文の投稿者へ Data Submission Form を送付してもらうよう依頼の手紙を出しました。 1986-1987年 に故丸山教授が発送した依頼状のUpdateです。御協力頂く雑誌の編集部には深謝いたします。

論文の投稿者へは Data Submission Form が送付されると思いますが、御協力お願いします。

学術雑誌名	対応
Agricultural and Biological Chemistry	
Cell Structure and Function	
Development, Growth and Differentiation	accept時に著者に Form を送付
Microbiology and Immunology	accept時にpreprintをDDBJに送付
Japanese Journal of Cancer research: GANN	*
Japanese Journal of Genetics	accept時に著者に Form を送付
Plant and Cell Physiology	
The Journal of Biochemistry	
Zoological Science	*

*: 1986-1987年、故丸山教授が依頼した際、協力していただけるとの連絡があった雑誌

様

DNA データ収集に関する協力のお願い

近年、DNAデータは増加の一途をたどり、DNAデータベースは生物科学の幅広い研究分野で必要不可欠のものとなりました。　このような状況のもと、1982年 DNAデータバンクが米国(GenBank)、欧州(EMBL Data Library)に設立され、又日本でも、科学大国になりつつある日本もデータの利用ばかりでなくその提供の面でも相応の寄与が必要であろうとの考えから、1986年 DNA Data Bank of Japan(DDBJ)が日本におけるセンターとして国立遺伝学研究所遺伝情報分析研究室に設立されました。

現在までの DDBJ の主な活動は、1) DNA sequence の収集と入力： GenBank、EMBLと国際協力のもとでデータベースの共同構築　2) DNA (GenBank, EMBL, NBRF), Protein (PIR)データの配布　3) 遺伝研共同利用計算機を用いての DNAデータのオンライン利用のサポート：解析システムの開発　4) ニュースレターの発行等の広報活動、等です。勿論最大の業務はデータの収集管理です。

データの収集管理に関しましては、データバンクは入力能力を越える DNAデータの増加に直面し、データベース構築の国際協力の緊急性を認識し互いに密接な協力関係を築きつつあります。主なものを挙げますと、　1) 入力対象とする学術雑誌の分担　2) 学術雑誌への協力要請　3) Common DNA Data Submission Form （データ提出の共通書式）の作成、使用　4) 各データバンクの使用する Accession number の共同管理　5) GenBank, EMBL フォーマット（Feature & Sites レコード）の改良　6) 関係データベース（RDB）の共同構築、等です。

データの収集は専門知識を持った人間の多大の労力を必要とします。そのため収集入力がデータの報告に追い付けない状況です。そこで現在データバンク間で収集の分担をしております。収集入力システムの上から現時点では収集対象の学術雑誌を分担することが最適のため、DDBJ は日本で出版される学術雑誌を主に担当しデータを収集入力しております。また収集にあたってデータの報告者の協力を求めるのがデータの質を高めるためにも不可欠です。そのためデータバンク間で学術雑誌に協力を求めることになりました。各学術雑誌で状況が異なりますので協力の方法は少しづつ異なります。 DDBJ としては、DNA 配列データを含む論文の著者に Data Submission Form （フロッピーと印刷物）を送付しデータバンクへのデータの提供を著者に呼びかけていただけるよう、学術雑誌の編集部に協力をお願いしております。貴雑誌編集部におきましても、なにとぞ主旨を御理解の上ご協力下さいますようお願いいたします。つきましては下記まで御返事下さいますようお願い申しあげます。

参考のため、Data Submission Form （フロッピーと印刷物）を一部同封させていただきます。

1988年 11月 15日

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
宮澤　三造
Phone: 0559-75-0771 ext. 649

DNA 配列データ提出のお願い及び DDBJ, EMBL, GenBankへのデータ提出について

宮澤三造

近年、DNAデータは増加の一途をたどり、DNAデータベースは生物科学の幅広い研究分野で必要不可欠のものとなりました。 一方、データバンクは入力能力を越える DNAデータの増加に直面し、データベース構築の国際協力の緊急性を認識し互いに密接な協力関係を築きつつあります。主なものを挙げますと、 1) 入力対象とする学術雑誌の分担 2) 学術雑誌への協力要請 3) Common DNA Data Submission Form の作成、使用 4) 各データバンクの使用する Accession number の共同管理 5) GenBank, EMBL フォーマット (Feature & Sites レコード) の改良 6) 関係データベース (RDB) の共同構築、等です。

データの収集は専門知識を持った人間の多大の労力を必要とします。そのため収集入力がデータの報告に追い付けない状況です。そのためデータバンクは研究者の方々に Data Submission Form を送付しデータバンクへのDNA 配列データの提供を呼びかけております。データの報告者の協力がデータの質を高めるためにも不可欠です。なにとぞ主旨を御理解の上ご協力下さい。

データの収集に関しては、現在データバンク間で収集の分担をしております。 DDBJは日本の研究者が解析したデータを収集入力することを目指していますが、現時点では収集入力システムの上から収集対象の学術雑誌を分担することが最適のため、DDBJ は日本で出版される学術雑誌を主に担当しデータを収集入力しております。そのため、日本の研究者も外国雑誌に投稿することが普通ですので、EMBL, GenBank からデータ提出を依頼されることが多いと思います。 特に Nucleic Acid Research への投稿は現在 EMBL へのデータの提出が条件となっています。DDBJとしては、そのような場合も DDBJ にデータを提出してくださるようお願いしています。データは電子郵便で EMBL, GenBank に転送しますので時間のロスはありません。これは、日本の研究者に便利であるだけでなく、将来 DDBJがデータ提出先として認められるためにも役立ちます。どうぞ是非御協力お願い致します。なお、DDBJが担当入力したデータも勿論 GenBank, EMBL に DDBJ 経由で登録されます。

なおデータの提出は可能な限り計算機可読な形でお願いします。近年パーソナルコンピューターの利用が盛んですので、配列データを入力している方も多いと思いますからその点問題はないかと思います。媒体は

(1) 電子郵便によるもの
(2) フロッピーディスク：MS-DOS用フロッピー (5.25", 3.5") もしくは Macintosh 用
(3) 磁気テープ (9トラック)

でお願いいたします。 提出いただくのは

(1) Data Submission Form による配列データの注釈
(2) DNA 配列データ

DNA 配列データは Data Submission Form の最後に追加して下さい。 Data Submission Form がHardcopy版しかない場合にはフロッピーでお送りしますので巻末のソフトウエア申し込み書を用い請求して下さい。なおファイルを作成する際には、どんなプログラムでも読めるよう以下のことに御注意下さい。

(1) ファイルは単純なテキストファイルであること。
(2) 一行の長さは 80 字以下であること。行末は、必ず改行で終わること。

特に、パーソナルコンピューターでデータを作成する際ワードスター等を用いるときは、必ずnondocumentとして作成して下さい。documentとしてファイルを作成するとワードプロセッシング用の文字が挿入され、他のシステムに転送した時意味をなさなくなりますので御注意下さい。

一番便利なのは電子郵便によるデータサブミッションですが、日本では残念なことにまだ普及していません。DDBJの共同利用計算機は、現在電子郵便ネットワークは接続されています。データ提出等の目的のための電子郵便の利用は、利用登録をしなくても使用可能です。どなたでも利用できます。内線電話、外線電話を利用しDDBJの計算機にパーソナルコンピューターを接続し、ddbjnewsでlogin し、データを転送しそのファイルを電子郵便でDDBJ(DDBJsub@niguts.nig.junet), EMBL(EMBLsub@niguts.nig.junet)又はGenBank(GBsub@niguts.nig.junet)宛発送して下さい。DDBJは、そのメールを各データバンクへ転送いたします。 Data submission form も、オンライン用のものがDDBJ計算機から手に入ります。以下に参考のため DDBJ計算機に login し submission する手続きを示しておきます。詳しくは、利用の手引を参考にして下さい。

パーソナルコンピューターによる接続にはモデムが必要です。費用がかかりますし、又ファイル転送、メール発送等慣れが必要ですが、数時間で届きますので最善の方法です。方法等、利用の手引に詳しく載っていますが、不明な点はお問い合わせ下さい。またUNIX計算機の場合には JUNET に加入し電子郵便を直接送受信することが可能です。NEC PC9801でもハードディスクがあれば UNIX(PC-UX) を購入すれば可能です。UNIX計算機をお持ちの方は、0559-75-0771 内線 647 宮沢までご連絡下さい。

How to submit data to databank

1) Get a online submission form by using "getinfo".

```
% getinfo
...
item ? DDBJ_news*                # choose DDBJ-news
...
item ? data_submit               # choose data_submit
...
item ? ddbj-form                 # choose ddbj-form

page ...  : h                    # get help
...
page ...  : s ddbj-form          # save ddbj-form into "ddbj-form"

page ...  : q                    # quit
item ? q                         # quit getinfo
item ? q
%
% ls -li ddbj-form              # make sure there is a ddbj-form
```

2) Transfer ddbj-form into your PC; it is assumed that kermit is used at the PC.

```
% kermit
C-Kermit> server                 # enter server mode
...
(Type ctrl-] c. to get ms-kermit command mode.)
Kermit-MS> get
Remote Source File: ddbj-form
Local Destination File: ddbj.frm
...
Kermit-MS> finish
Kermit-MS> connect

C-Kermit> exit
%
```

3) Edit ddbj.frm at your PC; put DNA sequences and others at the end of form.

```
Your electronic mail address is
        your-loginname%niguts.nig.junet@vax2.nlm.nih.gov
if you have an account for the "niguts". Otherwise, leave
its field blank.

- Please be careful to make files readable by any program; especially
  if you make it in PC.
- Files must be simple text files; nondocument-open for Word Star.
- Each line must be shorter than 80 characters and ended by
  <CR> and/or <LF>.
- If you want to use floppy for data submission, please don't forget
  to format floppy disk compatibly with IBM-PC; see "ibm-pc_floppy".
- You may obtain a floppy diskette of submission form from the DDBJ.
```

4) Send ddbj.frm to the host computer; it is assumed that kermit is used at the PC.

```
% kermit
C-Kermit> server
...
(Type ctrl-] c. to get ms-kermit command mode.)
Kermit-MS> send ddbj.frm ddbj.frm
...
Kermit-MS> finish
Kermit-MS> connect

C-Kermit> exit
% conv -filter ddbj.frm          # remove ctrl-Z EOF marks.
% pg ddbj.frm   #make sure that it has correctly been transformed
```

5) Send ddbj.frm to an appropriate databank by electronic mail.

The data banks agreed to share journals that each data bank scans for data entry. So, if your data is published in one of those journals, please submit your data to the data bank that is in charge of that journal. "journal-list" shows what journals each data bank scans.

```
Mailing address (...@niguts.nig.junet):
  ddbjsub or DDBJsub    data submission to DDBJ
  emblsub or EMBLsub    data submission to EMBL
  gbsub or GBsub        data submission to GenBank

Example:
% mailx -s "sequence name submitted" emblsub $LOGNAME <ddbj.frm
% (wait 30 sec)
% mailx
...
? (type mail number to see the copy of the mail that you have sent to
   "emblsub".
...
? q
%
```

6) You will get an acknowledge from the databank within a few days by electronic mail.

6) Thanks for your cooperation.

第一回 DNA 配列データベースのための国際諮問委員会　会議報告

1988年 2 月 15 日 - 16 日

NIH, ベセスダ市、アメリカ合衆国

国立遺伝学研究所
遺伝情報分析研究室
宮澤三造

第一回 DNA 配列データベースのための国際諮問委員会が 1988年 2月15日から16日までアメリカ合衆国ベセスダ市にある NIHで開かれた。DNA Data Bank of Japan 担当者として参加を要請された。以下はその報告である。議事日程、参加者のリストは附属書類を参照下さい。

国際諮問委員会は、 1987年 2月25日 - 27日西ドイツのハイデルベルグで開かれた NIH/EMBL ワークショップ「分子生物学におけるデータベースの将来」で論議され提出された勧告に従い開催された。この勧告は、DNA データバンク、即ち欧州の EMBL Data Library、米国の GenBank、日本の DDBJ、の成功はこれら３つのデータバンクの緊密な協力関係如何に依るとの認識から出されたもので、この協力を推進するためワークショップは、このような国際協力を調整する目的で国際諮問委員会の設立を勧告したものである。

委員会のメンバーは、欧州、米国から各３名、日本から２名である。委員会のメンバー以外の会議の参加者は、データバンクを財政の面でサポートしている組織の代表者（NIGMSの Director、GenBank担当官と EMBLの Director-General）と各データバンクの代表者である。その他、データバンク活動に関係した諸機関の代表者、例えば ヒューマンジェノムプロジェクトに関係して米国エネルギー省の担当官、米国蛋白質データバンクの NIH 担当官等がオブザーバーとして参加した。

会議は、データバンク担当者の活動報告から始まった。この活動報告は勿論国際協力に焦点を合わせてなされた。以下その報告を簡単に述べるが、ここで報告された国際協力に関する計画は、 DDBJ も参加した 1987年 11月開かれたデータバンク担当者（各データバンクから数名）からなる会合で議論され、まとまった結論である。（この会合は毎年開かれ国際協力で行う共同作業について議論している。）

GenBank は、1987年これまでの NIH と GenBank (BBN-LANL)の間の契約の更新を迎え、新規契約は IntelliGenetics と Los Alamos Laboratory のグループが獲得したことを報告し、LANL はデータ収集を IntelliGenetics はデータ配布をオンラインによるデータ提供を受け持つ新体制について簡単に述べた。獲得した予算は、5年間で1700万ドルとのことである。

DNA 報告論文収集に関して、現在は雑誌をスキャンして収集しているがその完全性を保証する目的で文献データベースを利用する可能性に関して、その可能性はあるものの、DNA 論文検索にふさわしいキーワードがこれまでなくようやく最近追加されたため、その評価がまだ十分ではないとの報告がなされた。しかし、スキャン対象外の雑誌に報告されるDNA 論文の発見には、文献データベースを使用するとのことであった。後者に関しては DDBJ も同じことを考えていて、会議に出席のついでに、会議に出席していた国際的医学文献データベースである MEDLINE 関係者 (Dr. Benson, National Library of Medicine)に会い、協力を依頼した。彼は快く依頼に応じ、日本で出版される雑誌に関し DNA 論文を定

期的に検索しその結果を電子郵便で毎月送付してくれることになった。ついでながら御報告しておく。

また GenBank が中心となって作業を進めている、関係データベース(Relational Data Base) の構築に関し、作業が最終段階にあるとの報告がなされた。すなわち、最終的にはGenBank/EMBL/DDBJ で全く内容の同じ関係データベースを個々の所で構築し維持するという計画である。この新システムは、著者によるデータ入力にも適し、そのためのソフトウエア (パーソナルコンピューターの上で動くデータエントリーソフトウエア) を IntelliGenetics が作成する計画していることが報告された。この計画では、データバンク間でのデータ交換は迅速であることが要求され、現在データバンク間の電子郵便のネットワーク (BITNET-ARPANET-CSNET-JUNET)に替わるより高速で、信頼性の優れたネットワークが要求されることが報告された。GenBank は、衛星リンクが最適との意見を述べた。またエネルギー科学ネットワーク (Magunet Fusion Energy Network と High Energy Physics Network からなる Energy Science Network) を利用する案も出され、EMBLと DDBJはその可能性を調査することにした。日本では現在高エネルギー研が HEPNET に、名大プラズマ研がMFENETに接続されている。 DDBJ の共同利用計算機はこの何れともネットワーク可能である。両研究所にその可能性を打診するつもりである。

EMBL は、EMBL が中心となって進めている Feature Table の改良案について報告した。この改良案は、共通の Feature Table を使用しようとするもので、1986年 5月及び 1987年の 7月と 11月のデータバンクの会合で議論され、最終案がまとまりつつある段階である。これは、現在の Feature Table の不備を補うとともに、EMBL とGenBank フォーマットの機械的な相互変換が可能になり実質的に同じになるということが強調された。

またデータの著者による預託に関しては、同じdata submission form を データバンクで使用することが報告され、データバンク共同で完成させた form が披露された。
一方、学術雑誌に協力を求めることに関しては、EMBL とNucleic Acid Research との協定が報告された。 1988年 1月よりNucleic Acid Research へ論文を投稿するには、データを予め EMBL に提出せねばならない という協定である。 NAR のエデイターで国際諮問委員会のメンバーの欧州側代表の一人でもある Prof. R. Walker は現在まで全くトラブルが生じてないと述べた。彼によると、投稿者の 60 % は投稿前に EMBL から accession number を得ているとのことであった。 一定期間終了後、彼は NAR のエデイターとして他の学術雑誌のエデイターに手紙を出し、このようなシステムを採用しても何の問題もなく、データバンクに協力することを薦めてくれることになった。

DDBJ は 1986年 4 月正式発足以来の活動を報告し、少数ではあるが、半年おきに 2 度データをリリースしたことを述べた。またGenBankとは異なり、DDBJ はデータ収集とオンラインによるデータ提供、ソフトウエアの整備、開発等、GenBank における LANLと IntelliGenetics の両者の役割を果たさねばならないことを報告した。また人材の確保が難しい等を説明し、国際協力のもとで役割を分担することが DDBJにとって重要であることを報告した。また現在は日本で出版される雑誌を分担しているが、将来は日本の研究者のデータを受け持ちたい希望を述べた。勿論、そのためには、著者によるデータ入力の推進が望ましいことを強調した。その間、GenBankと EMBL 担当のデータはDDBJが窓口として取りまとめて提供する準備が完了していることを報告した。なお DDBJの担当者としての私の報告の詳細は、附属の原稿を参照して下さい。

このような報告が、相互協力に関するデータバンク担当者のパネルデイスカッションも

含め、会議一日目の午前中になされ、午後、1 時間30分程国際諮問委員会のメンバーだけの議論が持たれた。また、午後には質問に応える形で現状についてより詳しい報告がなされた。国際諮問委員会のメンバーだけの会議のまとめは 2 日目に勧告の形で報告され、それに関し数時間委員会のメンバーとデータバンク側との間で議論が持たれた。以下国際諮問委員会の考え及び勧告を述べる。

まず、データバンク間の協力（重複を避けるための雑誌の分担、accession numberの共同管理、common data submission form の採用等）は大いに評価された。一方、メンバーの関心の一つは、 2，3 年前に発表されたデータがいまだデータバンクに入ってないものが多々あるということであった。これは、データ入力の遅れの解消という観点から議論された。

国際諮問委員会のメンバーの感じた失望は、フォーマットに関し common data items の採用に向けてそれほど進展がなかった点である。メンバーは、 common data items の採用はファーマット変換に費やす労力を軽減する上で最重要と考えた。また利用者が一ケ所でEMBL，GenBankのデータベースを維持管理するに要する労力、計算機資源を軽減する意味で、データの同一性が求められた。これを実現するために、現在のデータベースに含まれる程詳しい情報は望まないから、必須な情報を含むフォーマットでデータを配布することが提案された。またこのようなフォーマットでデータ入力することは、データ入力作業を簡単にし、データ入力の増加をもたらすと主張された。これに関しデータバンク側は、既に common feature table が完成間近であることが主張されたが、委員会は提出された common feature table の最終案を見て、複雑すぎるとの印象を述べた。　またデータの注釈に関し、解釈が主観的な事柄までが含まれるのではないかとの考えが述べられた。

勧告は以下のようである。

1）データバンクは最低必要な common data items に関して一致すべきである。
2）データバンクは共通の feature table を採用すべきである。それにあたっては、データの過大な解釈を避けることを要請する。
3）データバンクはこれまで出版された文献にある DNA データで未入力のものを無くすべきである。
4）データベースは他のデータベースに関する相互参照に関して案をまとめるべきである。

最初の二つは、緊急に要請されるものであり、委員会としては、 6 ケ月以内に実施すべきである。

進展を見守るため、委員会は毎年開催されることを要請する。またしばしば報告をもとめるであろう。

データバンク担当者としては、現在データバンク間の協力は広範におよぶため、委員会のメンバーにその内容の詳細を理解してもらうには時間が足りないとの印象をもった。

なお、 "DNA Databases Monitored" (Science vol. 240, p. 375) に会議について報告されている。

From CZJ%nihcu.bitnet%cunyvm.cuny.edu%bionet-20.
arpa@RELAY.CS.NET Sat Feb 20 19:50:31 1988
Received: by nigsun.nig.junet (3.2/6.2Junet)
id AA26544; Sat, 20 Feb 88 19:50:30 JST
Received: by ccut.cc.u-tokyo.junet (5.51/6.3Junet-1.0/CSNET-JUNET)
id AA16888; Sat, 20 Feb 88 11:07:49 JST
Return-Path: <CZJ%nihcu.bitnet%cunyvm.cuny.edu%bionet-20.arpa@RELAY.CS.NET>
Message-Id: <8802200207.AA16888@ccut.cc.u-tokyo.junet>
Received: from relay.cs.net by RELAY.CS.NET id an25588; 19 Feb 88 21:01 EST
Received: from [128.92.192.5] by RELAY.CS.NET id aa10262; 19 Feb 88 21:04 EST
Received: from CUNYVM.CUNY.EDU by BIONET-20.ARPA with TCP;
Fri 19 Feb 88 17:55:41-PST
Received: from NIHCU.BITNET by CUNYVM.CUNY.EDU ; Fri, 19 Feb 88 20:57:21 EST
To: nucall@bionet-20.arpa
From: CZJ%NIHCU.BITNET@cunyvm.cuny.edu
Date: Fri, 19 Feb 88 20:57:08 EST
Received: from CSNet-Relay by utokyo-relay; 20 Feb 88 11:06:23-JST (Sat)
Status: RO

SUMMARY

INTERNATIONAL ADVISORY COMMITTEE FOR DNA SEQUENCE DATABASES

The International Advisory Committee for DNA Sequence Databases held its first meeting in Bethesda, Maryland, on February 15 and 16, 1988. The names of the advisors representing the United States, Europe, and Japan are attached.

The Advisory Committee was assembled following a recommendation of the NIH/EMBL Workshop held in Heidelberg, Federal Republic of Germany, February 25-27, 1987. The recommendation arose from a concern that the success of the DNA sequence databases, the EMBL DNA Sequence Data Library, GenBank in the United States, and the DNA Database of Japan, depends on the active collaboration of the three databases. To facilitate this collaboration, the Workshop recommended the establishment of this committee to offer coordinated advice to this crucial effort.

The meeting began with presentations from the database managers on efforts to date with a focus on collaborative aspects. Noteworthy has been success in several areas: division of the responsibility of scanning journals to avoid duplication of effort, the introduction of a scheme that prevents duplication of accession numbers, the adoption of a common data entry form.

One area that has not been resolved in the inclusion of common data items in the two original databases, GenBank and the EMBL DNA Data Library. This problem revolves around the difficulty in adopting a common features table, the table which is used to provide the annotation of the biological features of the sequence data. To resolve the differences, a workshop was held in May 1986 and followed by a meeting of representatives of GenBank, EMBL, and DDBJ in November 1987. The result of these meetings was a features table document developed with the purpose providing a comprehensive plan for annotation that solved many of the problems that both databases were experiencing with annotation. This document reached final stages just before the International Advisors Meeting.

Another concern of the Committee was that many sequences derived from articles published two or three years ago had yet to appear in any database.

Following discussion, it was the consensus of the International Advisory Committee that although the collaboration was going well with a number of achievements cited above, the committee was disappointed that further progress had not been made in achieving agreement on a common features table. Achieving such an agreement is critical to the continued success of the three databases because with the expected explosion in sequencing efforts, both in the United States and abroad, the databases will not have the luxury of spending manual effort to translate the entries from one format to another. To achieve this translation automatically requires the adoption of common data items. In addition, on the basis of a brief examination, the committee felt that the proposed features table scheme might prove too complicated for general use.

As a result of these discussions the committee made the following recommendations:

- The three databases should agree to a minimum set of common data items for each entry.

- There should be a common features table adopted by the three databases. The committee urged that the data within the database not be over interpreted.

- The databases should remove the backlog of sequence data appearing in the earlier literature.

- The databases should be working on plans to include pointers to other genetic databases.

Because of the critical nature of the first two goals, the Committee urged their implementation within six months.

To monitor progress, the committee agreed to meet annually and to request frequent progress reports.

FIRST INTERNATIONAL MEETING FOR DNA SEQUENCING DATABANKS

Monday, February 15

9:00 AM Introductory Remarks.....Dr. Soll
Dr. Kirschstein
Dr. Philipson
Dr. Uchida

Overview of Current Databank Operations

9:30 AM GenBank.....Dr. Kelly, Dr. Benton, Dr. Burks
10:00 AM EMBL Data Library.....Mr. Cameron
10:30 AM DNA DataBank of Japan....Dr. Miyazawa

11:00 AM Coffee Break

11:30 AM Panel Discussion of the GenBank, EMBL, DDBJ Collaboration.....
Dr. Peterson, moderator,
Dr. Burks, Mr. Cameron,
Dr. Miyazawa

12:30 PM Lunch at a nearby restaurant

2:00 PM Executive Session for Advisors

Committee Discussion on their Role as Advisors for the Databases

3:00 PM Coffee Break

3:15 PM Discussion of Specific Issues

-The need for common data items in the different databases.

-The need for rapid exchange of data among databanks.

-The advisability of databases proceeding with Phase II agreements.

-The creative use of curators to monitor completeness and to update entries.

-The role of the CODATA committee in coordinating databases.

-The ties between the DNA databanks and the protein databanks have increased. Should this trend increase?
-Development of interfaces for access to databases.

5:00 PM Adjourn

6:00 PM Dinner at a nearby restaurant

Tuesday, February 16

8:30 AM Continued Discussion of Specific Issues

10:30 AM Coffee Break

12:00 noon Lunch at a nearby restaurant

1:00 PM Continued Discussion of Specific Items

2:30 PM The meeting is closed.

Participants

1. Members of Advisory Committee

From Europe

Prof. Richard T. Walker — Molecular biologist, Editor of Nucleic Acid Research
Department of Chemistry of the University of Bermingham
England

Prof. Piotr Slonimski — Molecular biologist
CNRS
Centre de Genetique Moleculaire
France

Dr. Rolf Fritz — Computer scientist
Deutshes Institut fur Medizinische Dokumentation und Information
West Germany

From U.S.A

Prof. Dieter Soll — Molecular biophysicist
Department of Molecular Biophysics and Biochemistry
Yale University
New Haven

Dr. Michael Waterman — Applied mathematician: sequence analysis
University of Southern California
Los Angles

Dr. Michael Coombs — AI researcher: natural language
Computer Research Laboratory
New Mexico State University
Las Cruces

From Japan

Prof. Hisao Uchida — Molecular biologist
Teikyo University

Prof. Minoru Kanehisa — Biophysicist
Kyoto University

2. Representatives of granting organizations

Dr. Lennart Phillipson
Director-General
EMBL

Dr. Ruth L. Kirschstein
Director
National Institute of General Medical Sciences

Dr. James Casatt
Genbank project officer
National Institute of General Medical Sciences

Dr. Jane Perterson
Genbank project officer
National Institute of General Medical Sciences

3. Staff members of databanks

From EMBL Data Library

Dr. Graham Cameron
EMBL Data Library

From GenBank

Dr. Mike Kelly
Principal Investigator
IntelliGenetics, Inc.

Dr. Christian Burks
Co-Principal Investigator
Los Alamos National Laboratory

Dr. David Benton
GenBank Manager
IntelliGenetics, Inc.

Dr. Tom Marr
Computer scientist
Los Alamos National Laboratory

From DDBJ

Dr. Sanzo Miyazawa
National Institute of Health

OBSERVERS

.
.
.
.
.
.

For the "First International Meeting for DNA Sequencing Databanks" held at the NIGMS from Feb. 15 to 16, 1988.

Activity of the DNA Data Bank of Japan

Sanzo Miyazawa

DNA Data Bank of Japan
Laboratory of Genetic Information Analysis
Center for Genetic Information Research
National Institute of Genetics
Mishima, Shizuoka 411
Japan

E-mail: ddbj%nigsun.nig.junet@relay.cs.net

DDBJ staffs

Sanzo Miyazawa	Manager/Database administrator
Hidenori Hayashida	Scientific Reviewer
Motono Horie	Secretary

1. Introduction

DNA data bank of Japan was established in the National Institute of Genetics with grant from Japanese government in April, 1986. This would be a result of efforts of many people, especially Profs. H. Uchida, T. Ooi, and M. Kanehisa, who realised importance of the databank and its impact on the research of bioscience in Japan and also in the world, and then persuaded the government to establish the databank in Japan.

Support for the DDBJ is a commitment of the government and the grant for running the databank is not temporary but rather permanent. The National Institute of Genetics is fully responsible for running the databank. At present, this project consists of two full time faculty positions, some running budget, and computer facilities. This organization consisting of staffs, machine and building has been completed in April, 1987. Since this project started in the Institute, it had been directed by late Prof. Takeo Maruyama, who was supposed to be here but regrettably died last December.

There are two advisory committees for the DDBJ, one of which is a committee within the Institute and other is run independently of the Institute and chaired by Prof. Uchida. The Uchida committee consists of a wide range of scientists from basic to applied fields and from computer specialists to experimental molecular biologists. The function of both committees is to guide the Japanese project of the database which has just begun.

Under their advises, the DDBJ is operated by two scientists, a secretary and a few annotators who are all part-timers.

2. Tasks of DDBJ

Tasks of the DDBJ are

1) DNA data collection and data entry in collaboration with other databanks,
2) data distribution, including secondary distributions of GenBank and EMBL data in Japan,
3) to provide researchers on-line access to DNA databases and programming tools for sequence analysis, and
4) to publish newsletters for an advertisement of DDBJ activity.

In the case of GenBank, the LANL does data collection and the IntelliGenetics provides data distribution and on-line access to databases. In our case, DDBJ has both functions of the LANL and IntelliGenetics. Beside data collection, we redistribute GenBank, EMBL, and NBRF-PIR databases in Japan with permission from those databanks. In addition, We must develop and provide research tools for DNA and protein information analysis. To let people know such activity of the DDBJ, we have published newsletters several times. The newsletters contain articles that describe the state of international collaboration among databanks and others such as a EMBL-NIH workshop held at Heidelberg last February and a meeting for staff members of databanks held at the IntelliGenetics last November, and how to submit data to databanks as well as matters of interest specific for Japanese scientists such as available databases, how to access the DDBJ computer system and how to use the databases. We hope that the newsletters can serve the community of Japanese scientists to realize the international trend of databank activity. Such an advertisement of databank activity is also conveyed through on-line service of the DDBJ computer system; any information of our activity may be obtained by accessing the computer system.

In the following, I will briefly report about the present state of data collection and data input by the DDBJ.

3. Data collection

DNA data collection and data entry are a primary task of the DDBJ. Our data collection is carried out in collaboration with the GenBank and EMBL.

We started collecting DNA data in December, 1986. Before starting data input, we had a lot of discussion about what kind of data should be collected, and what format should be used, and so on. Speaking of data format, we thought that we should not use a new format but either one of the GenBank or EMBL format. We decided to use the GenBank format, because a format which had been used in Japan is almost the same as the GenBank format. About one year ago, we had an experience to input unannotated entries but no experience of making annotated entries. So we felt that we needed such an experience, and then we regarded the first a few months as a learning period to make annotated entries. In this period, we dealt with papers published in a wide variety of journals. As a result, the DDBJ release 1 included some entries which were already input by other databases, GenBank or EMBL.

When we started collecting DNA data, we did not have a computer system for data entry. The computer system became available in April, 1987. Since then, we have been making a data entry system. Because error checking programs became working, we released the first version of the DDBJ in July, 1987. The release 1 included only 66 entries and 108,970 bases. Half year later, January, 1988, we released the second version which included almost twice as much as data of the release 1, that is, 142 entries and 199,392 bases. The release 2 included files of journal index, accession number index, short directory, and data submission form as well as DNA data.

Figure 1. Release note

```
                DNA Data Bank of Japan

                Release 2,  January 1988

            142 loci, 199392 bases, 8943 lines

This data base may be copied and redistributed freely,
without advance permission,  provided that  this
statement is reproduced with each copy.

Prepared by:

Takeo Maruyama       General Manager
Sanzo Miyazava       Manager/Database Administrator
Hidenori Hayashida   Scientific Reviewer
Motono Horie         Secretary

Files included:

1) relnotes.txt: this note
2) ddbj.dna:  DNA data
3) ddbj.cds:  peptide coding sequences extracted from ddbj.dna
4) ddbj.pep:  peptide sequences translated from ddbj.cds
5) journal.idx:  journal index
6) accnum.idx:   accession number index
7) shortdir.idx: short directory
8) datasub.txt:  data submission form

The 3rd and 4th files above were generated by using
"seqext" and "peptir" programs made by Dr. Jim Fickett
in GenBank, Los Alamos National Laboratory.

Acknovledgements:

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

          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
  E-mail: ddbj%niguts.nig.junet@relay.cs.net
```

3-1. Data flow

Figure 2 shows data flow at the DDBJ. Data flow which we now obey is more primitive than data flows at the GenBank and EMBL. An unskilled person regularly scans predetermined journals, and takes photocopies of papers that include DNA sequences. One of DDBJ staffs who is working as a scientific reviewer looks over the papers and judges whether they should be processed or not. Annotation is done by qualified persons who are usually graduate students. One of undesirable things at this stage is that annotators work with coding sheets rather than entering data directly to computer by using editor. There are several reasons. Annotators are not familiar with computer and there is no support tool such as sequence editor available for them yet. Beside, some of annotators work at home or distant places. After annotation sheets are checked by a reviewer, they are dealt with by punchers; base sequences are checked by punching twice. Other portions of data are also checked for mistyping by annotators. After that, they are checked by using programs in respect to format, taxonomy, journal name, and start and stop codons and codon frame in coding sequences. Those programs for error checking are ones kindly provided by the GenBank; they were programmed by Dr. Jim Fickett. I would like to thank him and the GenBank for such a help.

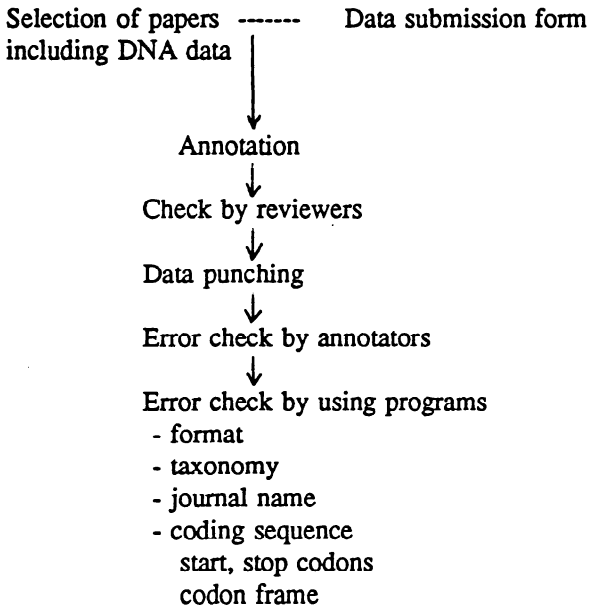


Figure 2. Data flow

3-2. Journals scanned

We have principally scanned journals published in Japan. Journal names scanned are listed in Table 1. In the last column, the first number means the number of entries and the second one within the parentheses is the number of papers, which are already processed and entered into the DDBJ database. If there are papers which are not processed yet, its number is shown at the end of line.

As expected, papers including DNA sequences do not appear in most journals published in Japan except a few journals such as J. Biochem. (Tokyo), Agricul. Biol. Chem., and Jpn. J. Genet. Even J. Biochem. (Tokyo) which published the most reports of DNA sequences in the Japanese journals included only 20-25 reports per year. The total number of papers including DNA sequences that were

published in the scanned Japanese journals below were only about 30-40 per year. According to the database search of MEDLINE that Prof. Uchida carried out, the total number of such papers are 17 with a keyword, "molecular sequence data", and 16 with "nucleotide sequence" in 1987. We are now planning to regularly scan such a few main journals, and use literature databases to search minor journals in which there are published only few reports of DNA sequences in a year. By the way, reports from Japanese research organizations were 148 of 1279 papers published in 1987 according to the BIOSIS Preview database. We collected about 140 entries last year. This number is nearly equal to the number of reports from Japanese organizations.

Because there are a small number of DNA sequences reported in Japanese journals, we have scanned a few other journals published outside of Japan. At present, we have charge of J. Gen. Virol. in the list. Other journals are officially scanned by the GenBank. A main obstacle to increase data entries is that it is not easy for us to keep many annotators. At present, we have only two annotators, and a reviewer all of who are part-timers. We will try to get as many annotators and reviewers as we can.

Table 1. Journals scanned by the DDBJ update: 01/31/88

		# entries	(# papers)		not entered
Journals published in Japan:					
Agricul Biol Chem	Vol. 50(01) - 50(12) 1986	3	(3)		
	Vol. 51(01) - 51(09) 1987	10	(9)		
Cell Struc Funct	Vol. 11(01) - 11(04) 1986	0	(0)		
	Vol. 12(01) - 12(04) 1987	0	(0)		
Chem Pharm Bull	Vol. 34(12) - 34(12) 1986	0	(0)		
	Vol. 35(01) - 35(06) 1987	0	(0)		
Devel Growth Diff	Vol. 28(01) - 28(06) 1986	0	(0)		
	Vol. 29(01) - 29(05) 1987	0	(0)		
J Biochem Tokyo	Vol.100(01) - 101(06) 1986	35	(22)		
	Vol.102(01) - 102(06) 1987	12	(2)	0	(15+4)
Jpn J Cancer Res	Vol. 77(01) - 77(12) 1986	0	(0)		
	Vol. 78(01) - 78(09) 1987	1	(1)		
Jpn J Genet	Vol. 61(01) - 61(06) 1986	9	(2)		
	Vol. 62(01) - 62(04) 1987	3	(3)		
Microbiol Immunol	We don't have this journal.				
Plant Cell Physiol	Vol. 28(01) - 28(05) 1987	1	(1)		
Zool Sci	Vol. 3(01) - 3(06) 1986	0	(0)		
	Vol. 4(01) - 4(05) 1987 (excl. 03)	0	(0)		
Journals published outside of Japan:					
J Gen Virol	Vol. 68(03) - 68(11) 1987	21	(13)	13	(12)
J Immunol	Vol.138(01) - 139(10) 1987	11	(5)	0	(40)
Immunogenetics	Vol. 25(01) - 26(06) 1987	0	(0)	0	(27)
	Vol. 27(01) - 27(01) 1988	0	(0)	0	(0)
J Gen Microbiol	Vol.133(01) - 133(11) 1987	0	(0)	0	(14)

Note: The last three journals are not officially taken charge of by the DDBJ.

3-3. Data entry system

Since the computer system became available in April, 1987, we have been making a data entry system. Usually a database system may consist of subsystems such as

1) data entry system,
2) retrieval system, and
3) data analysis system.

It would be ideal to manage all of the three systems by using a single management system. However, building such a system would take a time. We could not afford to choose such a way, because we already started data input. So we decided to make each system independently.

We built a data entry system by utilizing SCCS (Source Code Control System) available in our UNIX system. SCCS is a source management system with the following functions.

1) Version control
   A record is kept with each set of changes of what the changes are, why they were made, and who made them and when.
2) Exclusion control
   Only one person can modify data at a time.

Both are critical in data entry with more than one persons.

3-4. Quality control

As I already wrote, data are checked at several stages. Input of base sequence is checked by punching twice. Format, taxonomy, journal name, and start and stop codons and codon frame in coding sequences are checked by using programs made by Dr. Jim Fickett in the GenBank. However, sentences at reference, features, site and comments records, are not checked at all except by human review. We are planning to use a spelling-checker for this portion. Our experience of checking coding sequence regions indicates that non-coding sequence regions may include errors. Automatic checking of those regions by programs should be employed as well.

We hope that the quality of data entered by the DDBJ is comparable with that of data produced by the GenBank and EMBL.

3-5. On-line support for data submission

We are ready to help Japanese researchers to submit data to databanks including EMBL and GenBank. Our computer system joined the JUNET network for electronic mail and bulletin board. A special account is available for anyone to submit data to the databanks by electronic mail; of course, an on-line form for data submission may be obtained by using this special account. We will forward electronic mails to each databank with charge on the DDBJ, because overseas mails are charged and are not necessarily available to anyone in Japan. People can send DDBJ any data with any media. If journal on which that data is supposed to appear is not one which DDBJ has charge of, we will forward it to other databank by electronic mail. GenBank and EMBL may communicate with depositors through DDBJ.

4. Collaboration with other databanks, GenBank and EMBL

First of all, I would like to emphasize that international collaboration is critical for the DDBJ. Our databank was established only one year ago, and still at an prematured stage. We need know-how of

data entry which other databanks have acquired during data collection. My visit to the LANL last May with such a purpose bring us invaluable harvests. I learned how data flow is managed and what kind of programming tools are necessary for data entry and for quality control of data. However, there is one thing which I was embarrassed with. Because we intend to enter data in the GenBank format, we needed a detailed manual of the GenBank format for annotators and reviewers. As most of you realize, the present manual distributed with data was not sufficient for such a purpose. So one of purposes of my visit to LANL was to get such a manual. Of course, they had a good manual for annotation. What surprised me is that there were several discrepancies between their manual and the manual distributed. The manual distributed by BBN seemed to be out of date. They included examples of entries whose annotation is not completely right from the present stage of annotation. I think that this was caused by insufficient communication between them. This may not be a serious problem for usual users, but fatal for anyone who wants to enter data in the GenBank format. Keeping the same information at different multiple sites is not easy, but critical for us. One of such informations is taxonomical information. It is not necessarily needed for databanks to use the same taxonomical classification, although it may be helpful for users. However, it could save human resource needed for maintaining such information at each databank. DDBJ does not have such an expert. So it is critically important for us to get such information. Here I would like to ask GenBank and EMBL to help us to keep the same information as they have.

At present, a big project is planned by the databanks. It is to build a common relational database at different sites, that is, a distributed database. Keeping the same information among the databanks will be necessity in this project. We must devise a good way to do so. Including such a subject, the databanks, EMBL and GenBank, have been discussing what kind of collaborations are needed and how we can succeed them in order to accomplish this project. We, DDBJ, are ready to join such a collaboration between GenBank and EMBL. I would rather say that we have been in collaboration with them, since I attended the EMBL-GenBank meeting last November. I suppose that the contents of the collaboration among databanks will be reported by GenBank and EMBL. So I will stop my report here.

At the end, I would like to express my deep regret to late Prof. Takeo Maruyama who did not live to see DDBJ developed more completely.

NEWS FROM GenBank

Volume 1 Number 1 January 1988

GenBank Introduces Monthly Newsletter

By Alan Engelberg

Starting with this first issue, GenBank will regularly publish a newsletter. Users of GenBank share common interests, and the intention of the newsletter is to help draw the user community together.

We hope that the newsletter can serve the community in several ways. It will contain articles that describe how to use the data bank more effectively. The newsletter will post announcements of proposed and actual changes to GenBank and will serve as a forum, soliciting your reactions to the changes that are under consideration. Other articles will provide background information about the data bankand give users insight into the way GenBnk functions.

The staff of GenBank would very much like to hear from you and all GenBank users. We plan to publish a question and answer column in addition to letters and articles submitted by our readers. If you have a question but would rather not have it published, we will answer it, too. Please send your questions, letters, or articles by regular or electronic mail to:

Alan Engelberg, Editor
GenBank Newsletter
c/o IntelliGenetics, Inc.
700 East El Camino Real
Mountain View, California 94040

BITNET address:
ENGELBERG@BIONET-20

We hope this and future issues of the newsletter will help you in your research and encourage you to share your ideas with other members of the GenBank community. ◊

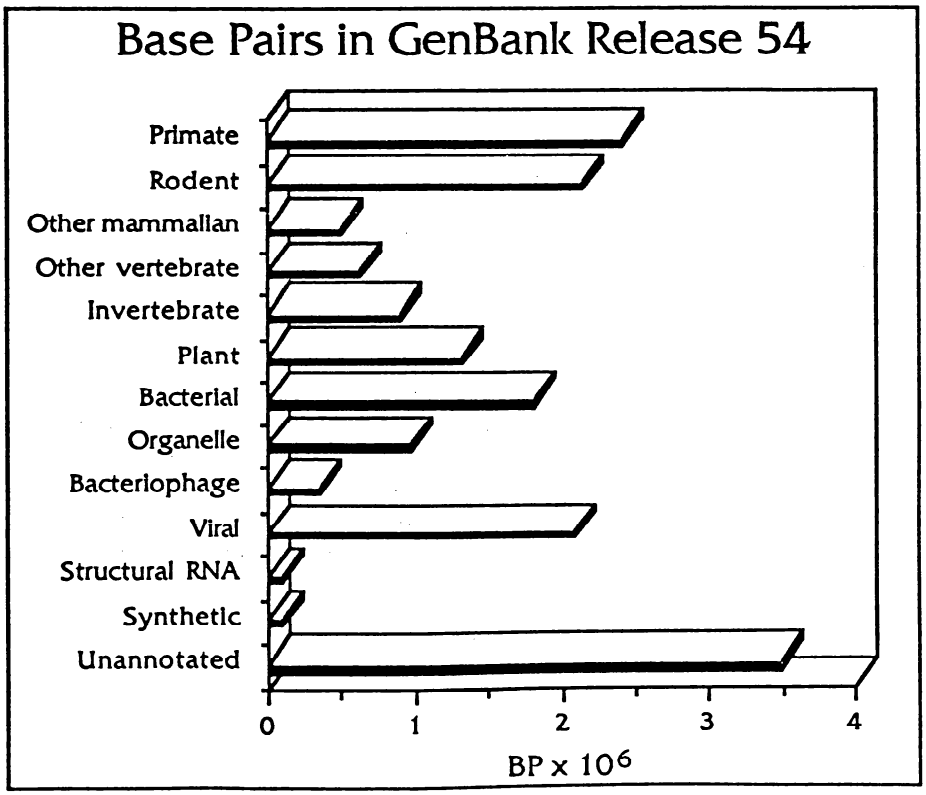
GenBank Release 54

By John Moore

GenBank Release 54 was distributed in January, 1988. It contains a total of 15,465 loci and 16,752,872 base pairs.

The accompanying graph shows the number of base pairs in each taxonomic category. The largest group in terms of base pairs is the primate sequences, which make up 14.4% of GenBank. The largest group based on number of loci is the rodent sequences, which has 14.5% of all the loci.

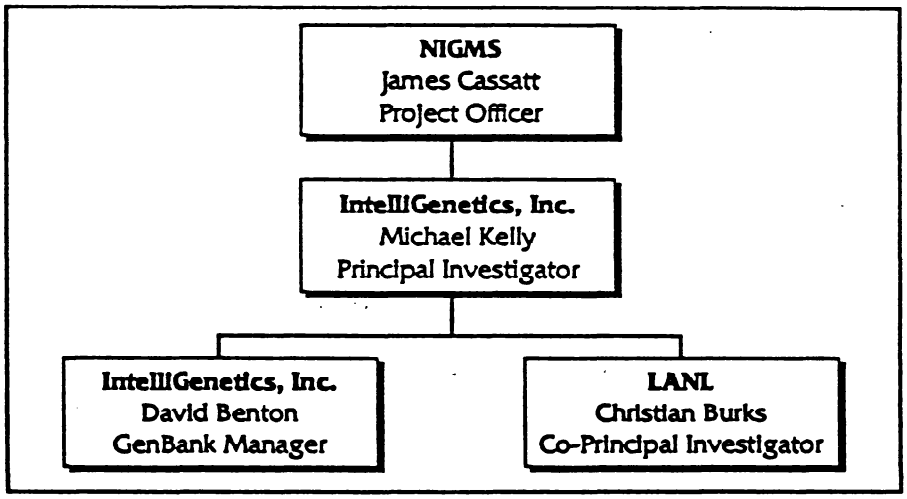
For comparison, Release 53 had 14,584 loci and 15,514,776 base pairs. One year ago, Release 47 had 10,485 loci and 10,388,356 base pairs. Release 54 thus represents an increase in base pairs of 8.0% over Release 53 and 61.3% over Release 47. ◊



The Structure of the New GenBank Contract

By Alan Engelberg and John Moore

On October 1, 1987, IntelliGenetics, Inc., in Mountain View, California, was awarded a five-year, $17 million contract to administer the GenBank Nucleic Acid Sequence Data Bank by the National Institute of General Medical Sciences (NIGMS) and other federal agencies. Los Alamos National Laboratory (LANL), which maintained the data bank during the first contract period (1982-1987), is still the subcontractor. The accompanying chart shows the relationship between the institutions responsible for Genbank.



NIGMS oversees contract performance

It became a division of the National Institutes of Health in 1958 and a full institute in January, 1963. NIGMS provides mostly extramural support to research and training in the basic biomedical sciences . Its offices in Bethesda, Maryland, hold a staff of only 156, making it the smallest of the institutes. However, its budget was $571 million in 1987, the fourth largest of all the institutes.

The Project Officer at the National Institute of General Medical Sciences (NIGMS) is James Cassatt, who supervises the project and is responsible for appointing Scientific Advisors to help set priorities and scientific policy in accordance with the needs of researchers. He also appoints Curators to help verify and integrate data.

LANL maintains and updates GenBank

It was founded in New Mexico during World War II and is operated by the University of California for the Department of Energy. LANL employs over ten thousand people and occupies more than 43 square miles; its operating costs in 1986 were $786 million. Most LANL projects involve research and development in defense and energy, including tunable lasers and fuel cells. Other projects explore basic physical sciences, mathematics, and computing. Biological research projects examine ways to make biocompatible materials for implants and to search for repetitive regions of DNA which may be used for chromosome pairing.

LANL has a Theoretical Division containing the Theoretical Biology and Biophysics Group, which has managed the data bank and has entered and updated all of its data since GenBank's inception in 1982. In 1978, Dr. Walter Goad, a member of this group, helped originate the sequence data base that became GenBank. Scientists in the group scan scientific journals for information and cooperate with the European Molecular Biology Laboratory Sequence Data Library and other gatherers of sequence data. The major work is in the annotation of sequences.

Christian Burks is the Co-Principal Investigator for GenBank and has the overall responsibility for GenBank standards, organization, hardware, and software. Thomas Marr and Chang-Shung Tung are other senior scientists involved in the development of molecular biology algorithms and in ways to improve the organization and use of the data bank.

IntelliGenetics distributes GenBank

It was founded in 1983 to write molecular biology software and distribute sequence data banks and molecular biology software to commercial and academic researchers. The company's founders were two computer scientists and two biologists from Stanford University who collaborated to develop a suite of programs for VAX, Sun, and DEC 2060 computers. IntelliGenetics also supplies molecular biology software for IBM and compatible microcomputers and has applied artificial intelligence techniques to the problem of cloning simulation and the management of plasmids. IntelliGenetics manages the NIH-sponsored BIONET Resource, an on-line computer network that enables more than 660 laboratories and 1700 scientists to search and analyze sequences.

The GenBank Manager at IntelliGenetics is David Benton. IntelliGenetics' primary role in the new contract is to distribute GenBank and develop software and systems to make it more accessible. The increase in funds in the new contract is partly intended to promote an on-line network of biologists using Genbank. Michael Kelly, President of IntelliGenetics and Principal Investigator for the GenBank contract, stated that the current contract is more than three times the dollar amount of the original one, reflecting the tremendous growth in the number of known sequences and the new needs generated by this growth. ◊

Date: March, 1988

Sanzo

Here's a listing of LANL's staff right now, which covers data collection, organization, and maintenance. You should get numbers for IG (data distribution and software for automation of collection) from dave benton.

Christian

Budget of GenBank:

The five year budget for the current contract is about $17,000,000, with about half going to IG and about half going to LANL.

task(s)	% FTE	total manpower
administration & strategic planning	50	
administration & strategic planning	50	1
strategic planning	5	
strategic planning	25	0.3
hardware system manager	100	1
programmer	100	
programmer	100	
programmer	50	
programmer	30	3.8
programmer & data flow management	100	
data flow management	100	
data flow management	100	3
annotation	100	
annotation	100	
annotation	100	
annotation	100	
annotation	100	5
sequence & citation entry	100	
sequence & citation entry	100	
sequence & citation entry	50	
sequence & citation entry	90	3.4
secretarial	100	
secretarial	5	1.05
		18.55

NOTE: At LANL we have now about >25 people working full and part-time, and adding up to about 15 FTEs; this will be going up over the next few months to about 35 people and 25 FTEs. I'm not as sure about IG; probably about 4-5 people now, adding up to 2-3 FTEs; but ask Dave Benton if you want full details.

We are in the process of expanding...by year's end we'll have added another 4-5 annotators, 1-2 programmers, and 1-2 administrative & strategic planning FTEs. (all these numbers are pretty rough, please keep in mind)

Date: March, 1988

Sanzo,

The best I can do for you quickly is the following table. I hope it is some help.

Graham.

EMBL Data Library Staff

Name	Room	Type of Work	%	Total manpower
Graham Cameron	348	Group Leader	100	1
David Hazledine	345	DB Admin.	100	1
Patrick Herde	347	Programming	100	
Shirley Jones	345	Programming	100	
Peter Stoehr	???	Programmer	100	
(replace S. Jones)	???	Programmer	100	4
Patricia Kahn	347	Biologist	100	1
Brigitte Boeckmann	347	Annotation	100	
Bernd Roechert	344	Annotation	100	
Guenter Stoesser	344	Annotation	100	
Michaela Sommerfeldt	Terminal	Annotation	20	
Baerbel Meissner	Terminal	Annotation	15	
Ruediger Rotfuchs	???	Annotation	50	
(to be appointed)	???	Annotation	50	4.35
Sylvie Karcher	346	Data Entry	100	
Karin Zojer	Terminal	Data Entry	50	
Charoula Christodoulou	Terminal	Data Entry	25	
Katrin Maste	Terminal	Data Entry	25	2
Tania Daskarolis	Terminal	Distribution	50	0.5
Amanda Lennon	346	Secretary	100	1
Rolf Apweiler	Terminal	SWISS-PROT	50	0.5

Total of 15.35 FTE's

Only 14.35 Simultaneous

第二回 DNA 配列データベースのための国際諮問委員会報告

1989年 2月 3日 - 4日

EMBL，西独、ハイデルベルグ

国立遺伝学研究所
遺伝情報分析研究室
宮沢三造

第二回国際諮問委員会が、欧州、米国から各 3人、日本から 2人（内田久雄教授、金久實教授）の計 8人の委員、予算当局を代表して米国 NIH、欧州生物研究所(EMBL)より各１人、遺伝研より瀬野教授の計 3人、オブザーバーとして、欧州共同体の担当官１人、データバンクを代表して米国 GenBankから 3人、欧州EMBLから 4人、日本DDBJから宮沢が出席して開かれた。会議はまず、第１回国際諮問委員会の勧告 4項目に関するデータバンクの対応について、データバンクスタッフの報告から始まり、委員は報告された内容に関し、おおむね了承した。またヒト及び酵母、大腸菌等のゲノム解析研究計画への係わり方について３機関からそれぞれ現状報告があった。一方、GenBank と EMBL の DNAデータバンクの運営費に比して DDBJ の運営費が一桁低額なため、委員会は日本の適当な機関への要望書提出を決定した。その他特に議論された項目をあげると、

1) DNAデータの注釈書式の更新についてその予定が報告された。
2) 研究者各人によるデータ入力サポート用のソフトウエアの開発の現状が報告された。
3) 過去報告されたデータでデータベースに未登録のものを無くす計画が報され承認された。
4) GenBank とEMBLデータベースを同一の内容のデータベースにするために解決すべき諸問題を議論された。
5) 現在データ入力は各データバンク間で協力分担している。よってデータ交換をスムーズに行うことは必須であり、そのためのシステムが議論された。委員はデータバンク側の処置に満足の意を表した。
6) ソ連アカデミーから国際 DNA データバンク活動への参加希望があったが、いくつかの理由から当面受け入れないこととした。

また第三回国際諮問委員会は遺伝研で1990年 3月16日-17日に開催することに決定した。

参加者としては、会議は国際協力推進というより米国、欧州の主張に終始し離反しているという印象を受けた。データバンク側としてはデータバン間での国際協力を側面から援助してもらうことを委員会に期待していたので、いささか心外であった。また委員会は、利用者の代表としてデータバンクの監督機関として機能することを目指しているとの印象を受けたが、DDBJ代表としては、DDBJは GenBankや EMBLとは違い契約として運営されているのではなく研究活動の一部であるので、研究者の自主性が保証されることが必要であるとの感を持った。

そのような委員会のはたすべき役割予期せぬ

Report of the Meeting of the International Advisory Committee with the Staff of the Databanks. Heidelberg Saturday 4th February 1989

In attendance

Advisors	Michael Coombs Rolf Fritz Minoru Kanehisa Piotr Slonimski Dieter Söll Hisao Uchida Richard Walker (Chairman) Mike Waterman
Funding Agencies	Jim Cassatt (NIH) Lennart Philipson (EMBL) Kanji Seno (National Institute of Genetics, Mishima)
Observers	Karl Heusler Benedictus Nieuwenhuis (Commission of the European Community)
Database Staff	David Benton (GenBank, IntelliGenetics) Christian Burks (GenBank, LANL) Graham Cameron (EMBL) Paul Gilna (GenBank, LANL) David Hazledine (EMBL) Patricia Kahn (EMBL) Sanzo Miyazawa (DDBJ) Peter Stoehr (EMBL)

Some of the International Advisors plus Jim Cassatt and Dick Nieuwenhuis had a preliminary discussion on the Friday evening. It was felt that this meeting was useful so that discussion could take place concerning the agenda and our approach to specific topics without the presence of Database staff.

"It is recommended that at all future meetings, the programme is arranged such that the International Advisors can meet separately for a few hours before and after the meeting and their travel plans should accommodate this." The representatives of the Funding Agencies could be invited to join part of these discussions.

One item for this meeting should always be a discussion of the report of the previous meeting. The chairman of that meeting should be prepared to explain which recommendations had been implemented and the reasons for any failures to do so. Some discussion took place centred on the excellent paper circulated by Michael Waterman, which raised the topic of the role of the International Advisors. It was felt that the key sentence in the document was the phrase 'one database' which despite internal differences in the way data is managed at the three sites, is how the project should appear to the outsider. We felt our main task was to make sure this goal was reached and then maintained.

We also raised some administrative problems and had the following recommendations...

"That in future, the agenda papers and accompanying documents should be circulated to participants at least one month in advance of the meeting and the tabling of documents at the meeting, should be discouraged as this leaves no

time for discussion." This means that agenda headings have to be made available to participants at least two months in advance of the meeting. A procedure for arrangements for future meetings needs to be agreed. Thus, someone has to take responsibility for assembling an agenda and circulating papers. Someone has to realise ahead of time that they are chairing the meeting and realise that a report has to be written and circulated. Also someone needs to ensure that the recommendations made are carried out. Although this can be discussed at the next meeting, some interim measures need to be taken at once and I suggest that T. Seno, as a general manager of the DNA research center of the National Institute of Genetics in which DDBJ is run should liase with Prof D Soell in preparing the agenda and other related subjects.

"As recommended last year, the International Advisors should receive reports during the year and not have to wait until the annual meeting. In particular reports on local meeting between Advisory Boards and Databank staff and the agenda should be circulated." This should not involve Databank staff in more work, we only want copies of what is already available. Professor Uchida has agreed to circulate reports on Japanese local meetings.

So that decisions taken at Databank staff meetings do not run contrary to suggestions of the IAB, the agenda for such meetings should be circulated to the entire IAB at least 14 days before the meeting and should invite comments.

"It is recommended that there is more overlap between the local American Advisory Board and the membership of the IAB. On several occasions it appeared that Databank staff were being given conflicting advice." "The annual IAB meeting should have more time allocated to it in future and should be spread over two days. This can no be doubt be combined into two full days with meetings of the IAB alone as recommended above.

"At the meeting on Friday evening and then again on the Saturday it was felt that now so much had been achieved, "The IAB should publicise these achievements". Suggestions ranged from notes sent to Science and Nature, to letters to the head of NIH .

Dieter Soell is dealing with this, particularly as far as NIH is concerned. In view of comments in a later paragraph (1a) any publicity to the outside world might wait for a year but this should appear on next year's agenda and plans then be made to do something.
From the papers circulated with the agenda, it was clear that much had been achieved since last year and the Databank staff are to be congratulated on the progress made and are exhorted to continue their efforts to complete the unification process. Apart from anything else a generally-perceived efficient, unified and world-wide Database will enable all concerned in the present exercise to attempt to repel the various current and planned attempts to hi-jack the more interesting parts.

The participants of the meeting agreed that the role of the IAB was primarily to achieve the position of "one database" and hence "to ensure that all nucleotide sequence data is in the public domain where it is accurately, immediately and universally available so that current and future important questions in Molecular Biology and related sciences can be addressed."

In practical terms this means that the IAB needs to ensure continuing database collaboration and funding for:

- Collecting
- Annotation
- StorageExchange
- Release

of data in the most rapid, complete and efficient manner possible.

The meeting started with a brief description of the funding of the three sites. The change from last year is the substantial commitment for two years by the EEC for the work currently at EMBL and for future continuity some negotiations for provision for 1992 onwards will need to take place immediately. The Japanese situation is not so fortunate as it apparently is difficult to get contract money and everything is regarded as research. It is also difficult for the Japanese representatives to get travel funds for meetings of Databanks. This means that the money available and therefore the resources, in Japan is an order of magnitude less than in the other two sites and maybe the IAB could send a letter to a suitably-placed person in Japan - either governmental or perhaps some Industrial Foundation. This is being investigated by Dieter Soell. We understand that the current GenBank contract ends in September 1992. It is hoped that the IAB will be kept fully informed concerning the planning of future funding for all the sites.

Points from the formal agenda

1(a) Feature Table.

We were told that the contents of this had been agreed in September and that an IBM PC Submission Program will be available in April. This will need publicity but the consensus is that it would be best left for a year until it is certain everything is working as expected. We had a discussion on gene naming which is the authors' prerogative but can (and does) cause confusion and a call from PS to include a form of words (which I have lost!) on the Data Submission Form concerning allelic genes (PS should let the Databank staff have the necessary information). The IAB is informed that the new Feature Table format is likely to appear in releases from September 1989 onwards.

The IAB views positively the fact that the databases are being developed in ways which allow the representation of synonyms in the technical vocabularies used.

1(b) Minimum Common Data Set.

This miraculously appeared at the meeting. Its appearance was greeted with relief and some confusion as there was some idea that it would be possible for the Databanks to provide just this material for scientists and that this was its main purpose. However, the report of the previous meeting made it quite clear that the primary purpose of this MCDS was so that unification of the databases was possible, independent of whether or not agreement on the Feature Table could be achieved. In the event, the existence of the MCDS has been overtaken by events and should be regarded as the minimum necessary information which accompanies a sequence, even though even this involves some degree of interpretation.

1(c) Missing Sequences.

We were assured by David Benton that the problem will not exist in two years. A determined effort is currently being made to clear the backlog and procedures about to be set in place by NLM under a reciprocal arrangement of critically-defining the mesh terms, means that in future essentially all (which of course means not all) sequences will be covered. Since the IAB meeting, GenBank has begun the process of entering the 1% it can identify as being in EMBL and missing from GenBank (see 2a below).

1(d)

Paul Gilna gave an impressive description of efforts being made to link with other Databases; such links are also being pursued at EMBL. "IAB recommends that every effort should continue to be made to promote collaboration between relevant Databases".

2(a) Overlap Between GenBank and EMBL.

Analysis of shared accession numbers indicates how much data common to the databases was actually generated by routine data exchange and suggests that subscribers to only one database might miss as much as 10% of the data. However, analysis by shared citations gives the impression that as little as 1% of the data would be missed by such subscribers. This contradiction is probably a result of very old data being entered at both sites under different accession numbers.

2(b) Data Exchange Mechanisms.

As an aid to controlled data exchange, the notion of accession numbers as unique, unchanging identifiers for information in the databases was endorsed by the Advisors. A document clarifying the rules for assignment of accession numbers (for example, what to do when a sequence if extended by a later scientific report) is being written for use between the collaborating databases and might be used as a starting point for a statement to the scientific community about how the system works.

We were given to understand that GenBank is able to update EMBL on a minute-by-minute basis so that sequences given accession numbers in the USA are available on the EMBL Fileserver by E-mail. DDBJ is ready to use a similar system. We understand that, given the general accessibility of the EMBL fileserver, Genbank has no immediate plans to offer a comparable service. The fileserver is obviously very much appreciated by the scientific community and its existence ought to be highlighted in any publicity material as one obvious positive gain from collaboration between Databanks and also journals. We also understand that once all three databases have their RDBMSs in place, have adopted a common set of data items to be exchanged, have adopted common definitions of and standardised vocabulary for those data items, have adopted a common format for exchange of data and have overcome the current shortcomings of currently available network connections, rapid exchange of data would provide no problem (!) and "all concerned are urged to reach this state as soon as possible".

"DNA データバンク国際協力のための実務協議" 参加報告書

国立遺伝学研究所
遺伝情報分析研究室
宮澤三造

1988年 7月と 9月の二度にわたりDNA データバンク国際協力のための実務協議が開かれ参加したので報告する。

1) New Feature Table共同作成に関する実務協議

会議は 1988年 7月 4日 - 9日まで西ドイツのハイデルベルグで DDBJ, EMBL, GenBankの Feature Table 担当者約 12 人が集まって開かれた。 DDBJからは宮澤三造が参加した。会議の目的は、1988年 2月開かれた第一回データバンクのための国際諮問委員会で出された勧告「データバンクは共通のfeature table（DNAデータ注釈項目）を6月以内に採用すべきである」に従い、DNAデータ注釈項目に関し議論することである。新データ注釈項目作成の協同作業は 2年前に発足して以来議論を重ね、この会議で最終案に関し合意ができるよう 2月以来電子郵便を用いて議論を深めてきた。そのため会議期間内に何とか一致を見いだし、最終案の作成を急いでいる段階である。今後の予定は 9月に新データ注釈項目のマニュアル（definition manual）を公表し、1989年の末ごろまでに新データ注釈項目を採用することを計画している。 新データ注釈項目案は共同学術雑誌に公表すると共にデータベースを配布している研究者へは直接通知されるであろう。参考のためにdefinition manual抜粋を付与する。

新データ注釈項目案採用までには以下のような作業が必要とされる。
1) 定義マニュアル(definition manual)の作成
2) 既データにおけるデータ注釈項目の変換
3) 新データ注釈項目によるデータ入力支援ソフトウエアの開発
4) データ注釈マニュアルの作成
5) データ注釈者の教育
6) ユーザーマニュアルの作成
このうち、2と3は容易ではなく時間がかかる作業である

2) DNAデータバンク定例会議参加報告書

会議は 1988年 9月 5日から 15日の間西ドイツのハイデルベルグで DDBJ, EMBL, GenBankの関係者総計 18 人が集まって開かれた。DDBJからは宮澤三造と林田秀宜が参加した。現在データバンクは入力能力を越える DNAデータの増加に直面し、データベース構築の国際協力の緊急性を認識し、互いに密接な協力のもと共同でデータベースを構築する計画を遂行している。このような研究計画においてはデータバンク担当者間の協議が欠かせず、最低年一回担当者による実務協議が行われている。参加者は、各データバンクで運営にたずさわっているデータベース管理者、データベースソフトウエア専門家、生物学者等である。今回の主な議題は

1）関係データベースデザイン
2）データ交換のためのフォーマット
3）分散データベースの同一性をいかにして保つか?
4）データ配布用データフォーマットの統一の可能性
5）データ入力支援ソフトの仕様
6）CD-ROM フォーマット
7）Curator システムの採用と問題点
8）データ収集の完全性をより高めるには? 等である。

最大の議題は次世代データベースとして計画している関係データベースデザインを詳しく検討することであった。しかし、分散データベースの同一性をいかにして保つか、データ交換の方法、等いくつか未解決の問題が残された。またデータ交換は初期の時点で 0.5 MB/日、数年後には 5 MB/日が予想されるためデータバンク専用の高速回線の必要性が指摘された。

会議ではまた以下のような事柄が報告された。

1）GenBank CD-ROM用のデータフォーマットに関するワークショップがサンフランシスコで1988年10月開催。CD-ROMの最初のリリースは1989年 4月に予定されている。EMBLは EMBL CD-ROMの最初のリリースを 1989年末に予定している。

2）GenBank は curator システムの採用に向けて準備

3）研究者自身によるデータ入力支援ソフトウエアのテスト版（IBM/PC用）が IntelliGeneticsにより1988年11月に完成予定。LANLはポータブルなものを開発している。 DDBJ、EMBLでもテストする予定。 DDBJはIntelliGenetics開発のIBM/PC用ソフトウエアをNEC PC用に移植する計画である。

4）GenBankはデータ収集を完全にするための予算（人件費：2名、他）を得た。

5）ソ連、インドがデータバンク設立に興味を抱いている。

諸外国の研究の現状

新 DNAデータ注釈項目作成は関係データベース構築と不可分の関係にある。新DNAデータ注釈項目の作成と関係データベース構築は、近年の分子生物学の進展に伴う新事実を取り入れると共に今後予想される人遺伝子解析等の大規模 DNAデータ解析の進展に追随できるようとの考えから第 2世代のDNAデータベースとして企画された。

GenBank担当者は関係データベースの完全な構築には18人年かかると見積っている。GenBank, EMBL 共に強く推進しており、ソフトウエア担当者の全て（GenBank 14人年、 EMBL 4人年）をこの計画に投入している。特にGenBankは人遺伝子解析には欠かせないとの立場から強く推進しており、関係データベースへの既データベースの変換をほぼ完了している。EMBLは遅れているものの1年後にはデータベースとしては完全ではないが一応移行可能な段階に到達することが予定されている。一方日本は、この計画を半年前に知ったためもあり、関係データベース管理プログラムを入手する段階である。またスタッフも 2 人と不足しており、関係データベース構築に関し GenBankや EMBLと同レベルのソフトウエア開発をすることは不可能に近い。その大部分をGenBankとEMBLから導入せざるを得ないであろう。計画の推進が強く望まれる。

参考とすべき点

最近、データ収集及びデータ管理を受け持っているロスアラモス国立研究所におけるGenBank関係の上級スタッフは全て人遺伝子解析計画(human genome project)に移行し、GenBankはその一部になった。これは、現在のデータバンク活動をこれまでの単なる延長ではなく、人遺伝子解析に向けてデータ管理、データ解析のための研究を目指しているからである。配列解析の実験技術の進歩から、最近ではファージの全遺伝子配列のような短いものばかりではなく、大腸菌（日本）、イースト染色体（欧州）、更には人間の全塩基配列までも解析しようとする計画（米国）が発足しつつある。遺伝子配列の長さが一桁、二桁あがる時、新しい実験技術が必要とされるのと同様に、データ管理、データ解析に関しても新しい方法が必須でありそのための研究が不可欠であろう。

一方EMBLでは欧州全域のサポートを強化すべく欧州生物科学ネットワークを推進し、欧州全域に散らばったプロジェクト参加研究組織の有機的な連携を計画中である。関係データベース構築も含みこの新計画では現在の16人相当のスタッフの倍増が見込まれている。（LANLでは25人相当のスタッフが現在プロジェクトに参加している。）このEMBLのプロジェクトは計算機生物学グループ（10-15人のスタッフからなるデータベース、DNA-蛋白質配列解析、蛋白質デザインなどを一括して研究する理論グループ）の活動の一つとして捉えられている。一方米国は生物学における計算機利用の促進を目的として国立ガン研(NCI)にAdvanced Supercomputing Laboratoryを 1985年に設立し、このような境界領域研究を推進している。現在、日本では計算機生物学の分野は研究者も少なく立ち遅れておりその充実が望まれる。

Collaborative Meeting
DDBJ
The EMBLData Library
GenBank®

EMBL, Heidelberg
5-15 September 1988

Report

I. Participants

Bob Abarbanel	(GenBank Consultant)	5-10th September
David Benton	(GenBank:IntelliGenetics)	5-14th September
Brigitte Boeckmann	(EMBL)	5-14th September
Christian Burks	(GenBank:LANL)	5-14th September
Graham Cameron	(EMBL)	5-14th September
Michael Cinkosky	(GenBank:LANL)	5-14th September
H. Hayashida	(DDBJ)	5-14th September
David Hazledine	(EMBL)	5-14th September
Patrick Herde	(EMBL)	5-14th September
Patricia Kahn	(EMBL)	5-14th September
Mike Kelly	(GenBank:IntelliGenetics)	5-10th September
Tom Marr	(GenBank:LANL)	5-14th September
Sanzo Miyazawa	(DDBJ)	7-14th September
Jane Peterson	(NIH)	-10th September
Peter Stoehr	(EMBL)	5-14th September
Günter Stößer	(EMBL)	5-14th September
Bernd Röchert	(EMBL)	5-14th September
Laurie Tomlinson	(GenBank:LANL)	5-14th September

Amanda Lennon Organising Secretary

II. Agenda

	Session 1 09:00-10:30	Session 2 11:00-12:30	Session 3 14:00-15:30	Session 4 16:00-17:30
5 September	*G. Cameron* Update on situation and plans at DDBJ, EMBL and GenBank	*G. Cameron* — contd. —	*G. Cameron* International Advisors meetings	*C. Burks* Existing Collaborative Tasks and Models
6 September	*T. Marr* Relationships between Data Sub. Form, Feature Table, Transactions, Distribution Format	*T. Marr* — contd. —	*G. Cameron* Identity of Database content	*G. Cameron* — contd.—
7 September	*P. Kahn* Journal Interactions, Submission Schemes and Forms	*P. Kahn* — contd. —	*L. Tomlinson* Data Flow, Entry Software	*L. Tomlinson* — contd. —
8 September	*D. Benton* Feature Table	*D. Benton* — contd. —	*D. Benton* — contd. —	*W. Ansorge* DNA Sequencing Workstation developments at EMBL
9 September	*M. Cinkosky* RelationalSchema	*M. Cinkosky* — contd.—	*D. Hazledine* TransactionProtocol	*D. Hazledine* — contd. —
12 September	*C. Burks* Curators and Outside Experts	*G. Cameron* The Two Formats Issue	*G. Cameron* Identity of Database content	*G. Cameron* — contd.—
13 September	*D. Hazledine* Source and definition fields	*C. Burks* Missing Sequences	*C. Burks* — contd. —	
14 September	*D. Benton* CD-ROM	*G. Cameron* Round Up andDates for Next Meeting		

(Chairmen shown in *italics*)

III. Reports

1. Update on plans at DDBJ, EMBL and GenBank®

DDBJ

- Our tasks are
  1) data collection
  2) to provide on-line access to various databases including protein sequence and structure database.
  3) to develop tools to analyze sequences.

- 2 full time employees (1.5 FTE) who are responsible for management and developing software. Part timer (FTE): 0.8 secretary, 0.2 reviewer, and 0.5 annotator

- We publish at least one newsletter in a year; two issues of about 40 pages were published this year.

- We use our computer system
  - to inform our activity to users as well as other information.
  - to provide a way for people to get online submission forms and submit data to data banks There are special mailing lists for EMBL (embl and emblsub) and GenBank (genbank and gbsub).

- We are very often asked by people of how they can submit data to EMBL or GenBank. We advertise how to make IBM compatible floppies and how to use E-mails. We will play a role of a gateway in data flow from Japan to EMBL and GenBank.

- We are in the phase I of "journal interaction". Prof. Maruyama got positive answers from most of Journal editors. At present, we are planning to ask Journals to send authors floppies of online submission form.

- We submitted a grant proposal in Aril:
  - one ph.D, one technician, and one part-timers; however, we could not get more than one post. We certainly need more people, but it is almost impossible to request more staffs in an usual grand proposal; we must find a way to get more people.
    (Only a way which we can take is to force researchers to use author-in programs.)
  - travel expenses for staffs and international advisors to attend the annual meeting; this seems to be rejected. So, we must find another way to get fund.
  - cost to buy the software of RDBMS.
  - cost to lease a packet communication line from Mishima to Tokyo to join BITNET or ESNET. In the case of ESNET, we need the permission from DOE and the High energy institute at Tsukuba to share an international communication line. Although priority is given for ESNET, it may be difficult to share the international line with the High Energy Institute; we do not have budget to share the cost.
  - cost for renting telephone lines and domestic network which are necessary to provide on-line access to researchers.

- We will buy the RDBMS by next March. We need strong supports from EMBL and GenBank in respect to softwares.

- There is a E. coli. genome sequencing project in Japan. DDBJ will take care of the sequence data analysed in that project.

EMBL

- The installation of the EMBL Data Library in the Oracle relational database management system is progressing, and remains a high priority. Changes, updates and improvements to the database are dependent on this.

- The present group is about 20 people working hours equivalent to about 16 full time employees.

- Expansion is being sought to (1) support the present operation to a plausible level, (2) research therequirements a new generation of nucleotide sequence databases to support large scale sequencing developments, (3) provide more comprehensive user support, (4) set up networks to nodes in various European nations who will provide local database support. Approval of all the current plans would result in a near doubling of the present group.

- Collaborations with various European publishers and database suppliers promise better future control of the completeness of the database. Derwent Publications is supplying sequence bearing patent data, and discussions are under way with Elsevier Science Publishers with a view to helping locate sequences in the scientific literature while scanning for their Excerpta Medica abstracts. This latter is similar to the GenBank® work with NLM's MedLine and should be developed so as to avoid duplication of effort.

- EMBL is pursuing the goal of researcher responsibility for database content in three ways: (1) directdata submission schemes, (2) "delegating" detailed annotation to experts in the field (Eukaryotic promoters are annotated by Bucher and Trifonov at the Weizmann Institute), (3) invoking the support of experts in dealing with special classes of entry (e.g., AIDS sequences).

- The printed detailed directories which EMBL used to produce with each release have been dropped, but effort is being dedicated to improving the machine readable indices.

- Progress is good on the new building at EMBL which will house, among other things, the Data Library

GenBank®

- GenBank is aproject of NIH, administered as a primary contract to IntelliGenetics, with a subcontract to the Los Alamos National Laboratory (LANL).

- At LANL the GenBank project exists in the context of a number of other genetic sequence related activities, among which are: (1) the Human Retrovirus and AIDS sequence database headed by Gerry Myers, (2) the Human Genome Information Resource, (3) the LiMB database of sources of information for molecularbiology.

- In addition to the "main" GenBank contract, there will be subsidiary support for activities in the areas of (1) work aimed at locating and including any previously unentered sequences, (2) extending the curator system, (3) providing the database on CD-ROM.

- The present LANL group is about 35 people working hours equivalent to about 25 full time employees. This work is divided into three main areas: (1) Data flow headed by Laurie Tomlinson, (2) Hardware and Software headed by Tom Marr, and (3) Annotation, for which a head will soon be appointed.

- Implementation of the database in the relational model is a high priority at LANL. An initial pass at conversion will occur soon.

- Other areas of emphasis include: (1) the annotation of the unannotated entries in the collection, (2)work with journals to improve flow of data into the database, and(3) identification of sequence bearing literature in collaboration with the national Library of Medicine.

- At IntelliGenetics the GenBank work exists in the context of a number of information technology projects for the biosciences.
- IntelliGenetics is working on improvements in data distribution including: (1) a greater range of tape formats, (2) CD-ROM, (3) online access to the data.
- Investigator entry software is being developed at IntelliGenetics.
- The GenBank Release Notes are being improved to more accurately and completely document the database.
- In October IntelliGenetics is hosting a Software Developers Workshop to explore the requirements of this group, communicate to them an overview of various aspects of the restructuring of GenBank, and to elicit feedback on some future formatting plans.

2. Existing collaborative tasks and models

The following is a summary of the attempt to structure and describe the collaborative tasks undertaken by the joint nucleotide sequence databases.

Though there have been and continue to be many tasks confronting us that are best addressed collaboratively, and though there are ofcourse many instances where we have been quite successful in working together to solve a particular problem, there have also been instances where our efforts have fallen short of either efficiency or fruition.

There are no doubt many reasons why we have fallen short in these instances. For example, we have often drifted or leapt into the collaborative mode without considering whether or notcollaboration was necessary or desirable in that particular instance. And whether or not collaboration was desirable or necessary, we have often not paused at the beginning of the collaboration to structurethat particular effort so that we have a clear understanding of which staff members are responsible for progress, how much feedback from other staff at either site is required, whether the feedback should (or should not) necessarily impinge on the iterative loops, and how we will recognize and measure progress.

It was therefore considered useful to have a discussionof:

- Different models of collaboration, independent of any specific tasks we are now or soon will be working on;
- For any of the specific tasks we are now or soon will be working on, which of these models would provide a useful framework for organizing the effort and optimizing the usefulness of the result.

The discussions in November 1987 resulted in a document that described various collaborative tasks and specified the modeof collaboration appropriate to each. The current meeting used this document as a starting point.

Discussion

The models document was evaluated as being a useful reference and catalyst; we will continue to maintain it.

The descriptions for the majority of tasks were not changed substantively. One task, the development of a product taxonomy, was downgraded in priority. Several new collaborativetasks were added to the list, including: standardization of journal name abbreviations in the three databases; development of adocument describing what reference files the databases are maintaining (or intend to maintain) in common and where the standard reference point is going to be; and development of a features table annotation standards guide.

Task List

The following collaborative tasks are thoseimportant enough at this point to list them and continue monitoring them inour discussions:

(a) Accession number domains.
(b) Journal split.
(c) Journal interactions, phase I (authors requested to contribute data).
(d) Journal interactions, phase II (authors required to contribute data).
(e) Hardcopy data submission forms.
(f) Computer-readable data submission forms.
(g) New features table format.
(h) Relational schema.
(i) Transaction protocol.
(j) Standards for inclusion/exclusion of sequence data.
(k) Author entry software.
(l) Organism Taxonomy file.
(m) Product Taxonomy file.
(n) Master list of current major collaborative tasks and approaches to addressing them.
(o) Journal Names.
(p) Master list of standard reference files for database software and protocols.
(q) Feature table annotation standards guide.

3. Relationships between: Data submission form, Feature table, Relational schema, Transaction protocol, Distribution format

The structure of the database is manifestedin many different ways, and it was felt worthwhile to spend some time discussing and defining the relationships between these different representations.

The **distribution format** is the way the data are seen by the user community. It consists of entries, where each entry contains information pertaining to a single sequence. An entry contains **sequence**and **annotation**. The **feature table** is the part of the annotation in which points andregions of significance are described. The rest of the annotation pertains to theentire entry. Up until recently the distribution formats have been the only representation of the information — all programs for data management have operated on it. The **data submission form** is a way of collecting the information needed to build an entry. Note that the form is a product of a collaboration between the three nucleotide and the three protein sequence databases (EMBL, GenBank®, DDBJ, PIR, MIPS, JIPID), and therefore gathers more information than is necessary for purely nucleotide sequence annotation.

The **relational schema** describes the tables the databases intend to use to store and manipulate the data in their relational database management systems. The relational schema is an efficient, butnot very human-readable way of storing all the information required for the database. It goes beyond the information included in the distribution format, or gathered by the submission form. It includes, for example, information about subscribers to the database and information about who at the databases entered what data.

The **transaction protocol** is the language the databases intend to use to communicate information to the database — both theirown local copy and copies of collaborators. The transaction protocol definition includes the **entity list.** This is a list of data items about which the databases can communicate in a standardised way.

The entity list must be agreed by the databases, and all three databases must implement a relational schema which can support the entities described. The data submission form gathers the information for these entities. The transaction protocol communicates this information to the databases, and the distribution format communicates it to the user. The feature tableis the part of the distribution format which gives information about points and regions on presented sequences. A subset of the entity list must be defined as the **minimal common data set** which should be supplied for all entries.

4. Feature Table

The common feature table format definition was adopted by the three cooperating data banks. Implementation of this feature table in public releases of the databases will occur as soon as possible after conversion of the databases from flat file to relational database management systems (for internal management of the data). The exact implementation schedule will be collaboratively determined. In the course of discussions of the feature table definition, a number of simplifications and additions were adopted:

- Feature nameswill be optional and will be represented as the value of a new qualifier, /label.
- Feature locations will be resolved to base positions (integers) whenever this does not impair readability.
- The qualifier /standard_name was added to allow each feature to be associated with the full name used for it in the scientific community and literature.

As an aid to implementation, a Feature TableAnnotation Standards Guide will be authored by the chief annotators from GenBank and the EMBL Data Library. A first draft of this guide is scheduled to beavailable in early 1989.

This feature table design depends on several controlledvocabularies in order to generate consistent feature representation. These vocabularies were identified, authoritative sources for the vocabularies wereidentified, and specific staff members were assigned to be maintainers of the internal database copies which will be common to the collaboration.

5. Data Acquisition: Journal interactions, Submission schemes and forms, Investigator entry software

We first discussed the "unassigned" (non-scanned) journals. GenBank reported that NLM has started sending them lists ofcitations to which the keywords "base sequence" or "molecular sequence data" have been assigned. Elsevier (Amsterdam) has been discussing doing similar work in collaboration with EMBL. GenBank agreed to send EMBL the NLM list on a monthly basis; when Elsevier begins producing a list EMBL will share it with GenBank. Since it might be desirable to contact the editors of those journals which occasionally publish sequence data (to inform them of our existence and our direct submission scheme), Laurie Tomlinson will divide up the journals from NLM's recent scan among EMBL, GenBank and DDBJ.

The data submission form was discussed briefly and there were no problems reported. Rather than updating it 4x/year it was felt that 2x/year would be sufficient. Patricia Kahn will discuss this with the protein sequence databases. LANL is interested in developing the on-line version of the form so that it is machine-processable. It was agreed that when they have a draft version they will distribute it to the other groups; if others want to adopt it, the task of producing the on-line form would be separated from that of producing the printed form (currently being done by Patricia) and would become the responsibility of Laurie Tomlinson.

Each database discussed its direct submission schemes and ongoing negotiations with journal editors. Both EMBL and GenBank are interested in getting journals to adopt a "phase II" scheme in which authors are required (rather than requested, as is the present practice with most journals) to submit their data to the database. EMBL and GenBank differ somewhat in terms of how they think such a scheme shouldbe set up. EMBL felt strongly that submission to the database should take place before the author submits his/her manuscript to the journal and that, as much as possible, we should try to get journals to adopt an identical scheme. GenBank reported that several journal editors they spoke with are adamant that submission to the database should be required only after acceptance of a manuscript. GenBank is less concerned about the existence of different schemes as long as they all converge once the data areactually submitted to the database. For the time being we agreed to disagree; as we continue to exchange information about our discussions with editors and our experience with phase II schemes, hopefully these differences will diminish.

IntelliGenetics is developing software to enable researchers to attach annotation to their own sequence data and to carry out some verification on the sequence and annotation. This software, which will have an extremely friendly user interface, will, in its first version, run on MS-DOS machines, but versions for othermachines will follow soon. Test versions should be available before the end ofthe year. EMBL and DDBJ intend to use this software rather than face the researcher with a multiplicity of interfaces.

6. Relational Schema

As has already been discussed, the relational schema specifies the tables in the RDBMS which will be used to store and manipulate the entities listed in the entity list, which is part of the transaction protocol document. EMBL and GenBankare both involved in building schemas to support the agreed entities and, although further work is required to finalise the entity list, schema designs are essentially complete. They will, of course, be modified in the light of the final entity list.

The core EMBL and GenBank schemas are the same, although there are more tables in the GenBank schema than that produced by EMBL. The tables in the schema handle 5 broad kinds of data:

- **Bibliographic data** — References (journals, papers, authors etc.) which relatemainly, but not exclusively to sequence data. Any database entity may have associated references.
- **Physical context data** — Data relating sequences to their biological sources (e.g, taxonomy, organisms, tissues).
- **Logical context data** — Defining features and functions of sequences and regions of sequences(e.g., genes).
- **Sequence Data** — Sequences are stored as presented in single scientific reports. That is, they are not merged.
- **Operational Data** — The various kinds of information needed to support our data processing operations. Not, in general, biological information, but rather information about things like what stage of processing a given entry is at.

7. TransactionProtocol

Substantial progress was made on the joint DDBJ/EMBL/GenBank transaction protocol. This comprises a set of entity definitions together with a syntax for specifying transactions which insert or update entities in the three databases. These entity definitions define a common set of data items which all three databases will be able to store and distribute. The transactions will be processable by computer programs, enabling us to automate and thus speed up data entry into each database. The transaction protocol is independent of any particular computer hardware, DBMS, database schema or programming language.

We envisage that it will be used for local data entry by each database, as a method for exchanging data between databases, and as a method bywhich researchers can submit data directly to the databases (e.g. by using dataentry software which generates transactions as its output).

8. Curators and Outside Experts

Direct acquisition of data from researchersis one way of tapping the expertise of the research community for the benefit of the databases. Both GenBank and EMBL are also pursuing other strategies. During thelast contract, GenBank had a system whereby "curators" who were experts in particular fields took some responsibility for helping with the data in their area. EMBL had similar, less formal arrangements within house researchers, and occasional help from outside experts.

GenBank is working on similar plans for the current contract, with more emphasis on direct help via computer networks from curators. EMBL has one extensive project where Bucher (Weizmann Institute) produces a promoter database which interfaces with the EMBL nucleotide sequence data library, and is exploring other collaborations, for example, with emphasis on particular organisms.

It was noted that (i) we should take care not to tread on each other's toes by avoiding (without prior arrangement) approaching potential curators in the "wrong" continent, and (ii) we shouldkeep each other informed as to the areas we aimto put under curatorial custody, to avoid overlaps. EMBL also stressed that the best such arrangements that they had seen in the past were temporally limited projects. The contribution of "tenured" curators tended to diminish over time.

There are slight differences in philosophy between GenBank and EMBL in the way in which we hope to draw on the expertise of outside experts. The GenBank curator system, and software development at LANL aims to equip curators to carry out direct work on the database. TheEMBL approach does not rule this out, but an explicit goal is to be able to accept help from people with differing levels of computer expertise, and in different computing environments. These differences seem harmless.

9. CD-ROM

CD-ROM is an obvious medium for future distribution of the Nucleotide Sequence data libraries. It has a high capacity, is very robust, and is cheap to produce and distribute. The devices required by users to read it are cheap, wellstandardised, and compatible with most kinds ofcomputer. Both EMBL and GenBank are planning soon to offer data on CD-ROM. (This is in addition to, not instead of, magnetic tapes.)

GenBank at IntelliGenetics has already done some detailed work on designing a CD-ROM format for the data. At the collaborative meeting it was decided to turn this work into a collaborative task, and work together to produce one CD-ROM format. We expect to agree on this format soon, but, at least at EMBL, there will be some delay before it can be implemented.

There are plans at EMBL to offer the present distribution format on CD-ROM, perhaps with some simple access software. Test systems are beingdeveloped with Phillips Du Pont Optical Company and Circle Information Systems. A stated requirement of the systems being tested that they should be able to use the collaboratively-developed, common CD-ROM format as soon as EMBL can provide data in this format.

10. Distribution Formats — the two formats issue

We often hear from users that the existenceof two distribution formats creates significant difficulties for them. Sometime was dedicated to discussing the nature of these difficulties, and whatwe might do to attempt to solve this problem.

- The contents of the two databases are not identical, and therefore users who wish to have as complete a data set as possible feel they have to take both databases. Format differences pose problems for people who take both databases.

  EMBL and GenBank both make releases every three months, and the databases interleave their releases, therefore, if we were both including all the data from the latest release of the other collection, GenBank would be most up-to-date half of the time, and EMBL the other half. In this

situation a user who took only one database would be sure that anything going into the other would reach him, but with about 6 weeks delay. This would be an improvement.

- The existence of two formats would remain a problem for users who wanted to take both collections to avoid delay, and for software developers who wanted their software to be able to process both databases.

- There is a common misconception that the failure of the databases to include all the data from the latest release of their collaborators stems fromthe differences in format. This is not the case. During the conversion data can fall into three categories:
  (a) Data present in the destination database — For example, GenBank entries converted by EMBL for which there are clear-cut corresponding entries the the EMBL collection. Such entries are not taken over in the conversion: we already have the data.
  (b) Data absent from the destination database — For example GenBank entries for which there are no corresponding data in the EMBL collection. These entries are taken over in the conversion.
  (c) Presence or absence unclear — Data which share a citation with entries in the destination database, butwhere there are other citations too,data which look as if they overlap with information already present. In short, data where our programs can't work out the relationship with data already present. Hitherto these data have not been taken over.

  It is the data in class (c) that pose conversion problems, to include them without generating duplicate information requires human intervention, and labour intensive. The problems in no way stem from the format differences.

- It is nonetheless clear that agreement on a common format would simplify the life of our users.

- Both groups are aware ofglaring inadequacies in both formats. There is little motivation to convertfrom one clearly inadequate format to another. Nonetheless, EMBL has seriously considered adopting the GenBank format. Earlier in the year this possibility was discussed with GenBank, and with our users. Unfortunately our users were overwhelmingly opposedto such a plan. This, coupled with our need to maintain compatibility between our nucleotide sequence collection and SWISS-PROT, our protein sequence collection, has led to a decision to continue to distribute the present EMBL format.

- The feature table has always been the major source of difficulty in attempting to standardise formats. The agreement on a common feature table is a major step forward, and could be a start towards a standard format.

- The databases see CD-ROMas a major distribution medium for the future, and **one common format** isbeing designed for that medium.

- The transaction protocolis a standardised language for talking to the databases.

11. Identity of Database Content

The perceived need of users to take both major databases in order to ensure completeness is a matter of concern. It is certainly the case at present that there is enough difference in the content of the two databases to cause a substantial population of users to find it necessary to use both databases. To have to deal with two databases in different formats is a severe headache. Approaches to ensuring that the contents of the two databases are the same were discussed atlength.

The contents of the databases differ in two senses:

- **Differentsequences are included.** GenBank includes sequences that EMBL doesn't include, and EMBL includes sequences that GenBank doesn't include. Sequences not taken over from other databases are either (i) those which were entered too late to be included in the conversion, or (ii) those which for some reason failed the conversion. The most typical reason for the latter is

that it looks as if is is already present in the receiving database, but our programs can't quite work it out.

- **The *same* are *differently described.*** In the early history of the databases, the division of labour was less well defined, and it was not unheard of for GenBank and EMBL to enter the same data. While, in general, the resulting sequences were the same, the details of annotation might differ — choosing keywords is a fairly subjective task.

  Even where the division of labour is working well, and a sequence is entered once and distributed to the collaborating databases, there remains the problem of propagation of updates. If an author writes to one database and asks for base 4235 to be changed from "a" to "t", we have to have mechanisms to ensure that that information is transmitted to the other groups. At present updates are propagated on a purely *ad hoc* basis — someone at one database communicates the need for the update to the other group.

Various approaches to ensuring that the samepopulation of entries are present in the different databases were discussed,ranging from throwing away one database and using the other as the starting point, to making one composite database from the sum of the existing databases and accepting, but slowly weeding out the resulting duplicates. Extracting a unique, but complete set of information from the two databases is a non-trivial task. Where references A, B, and C report contiguous sequences, they may appear in one database with an entry composed of the data from A and B merged, and in the other with B and C merged. Even where a simple one-to-one correspondence can be identified, it is not simply a matter of taking the "best" version, one database may have a more up-to-date version of the sequence, while the other has annotated the features of the sequence far more thoroughly. Teasing apart the parts of the two versions to produce a new composite version can be absurdly time consuming. Thus far this kind of work has been given a lower priority than developing systems to prevent this kindof divergence in the future. EMBL will, very soon, include on its release tapea file containing all the GenBank entries that didn't get through the conversion process because of some doubt as to whether they were duplicates of, or overlapped with entries already in the database. This file can be used by subscribers on a *caveat emptor* basis.

A major problem is that even supposing that we were starting with identical copies of the same database, there is no foolproof mechanism to ensure that they remain the same. The present mode of operation gives both groups the authority to carryout updates on the data. One can use the transaction protocol to inform the other group of updates in a way which allows theirprograms to carry out the same update. This doesn't completely solve the problem however. Problems arise if two groups simultaneously update the same record. E.g., suppose that one group carries out an update searching for a reference by an author "McBride" to change the start page from 126 to 1126 and **at the same time** theother group corrects the spelling of "McBride" to "MacBride" — the search on which the first update is based will fail for the second group, and without human intervention they won't be able to carry out the update. In the presentmode of operation, "at the same time" means within the same three monthly release cycle. Conventionally, databases prevent such clashes by "record locking systems" whereby data being updated are rendered inaccessible to other users until the update is complete.

EMBL's file server is an improvement on the quarterly release cycle, and GenBank has plans for continuous updating of their database, which should reduce the chance of clashes. The present network systems used to communicate between the databases are unlikely to support any workable record locking system. It is therefore planned to attempt to run the system without record locking in the first instance, but this decision may need to be reviewed.

12. Missing Sequences

It is often drawn to our attention by usersthat not all published sequences are included in the databases. Sequences in comparatively old literature are sometimes absent from either collection. To rectify this situation we need to (i) find all the old literature we have missed and include the sequences contained therein, and (ii) overhaul our data capture mechanisms to ensure that we are not continuing to fail to pick up missed literature. In both of these tasks the literature abstracting databases have

offered to help us — at GenBank the National Library of Medicine is helping through their MedLine database, on the European front Elsevier Science Publishers are helping through their EMBase (on-line Excerpta Medica) database. In locating old, missed literature the bibliographic databases can help by producing citation lists from queries designed to locate sequence bearing literature. In ensuring the completeness of future data capture they can help by explicitly tagging literature of potential interest to the databases as they scan the journals. Both GenBank and EMBL are exploring both approaches with their collaborators. Of course "hit lists" of potentially interesting, but missing citations will be exchanged and the workload dividedto avoid duplication in entering the relevant data.

Attempts to capture old literature via queries on existing data in the bibliographic databases meet with limited success. Much of what is located is irrelevant, and, when we compare with the citations included in the databases, much of what should beincluded is missed. It was felt that we would do well to announce our attemptsto gather old, missed sequences by publishing announcements in appropriate scientific journals and on electronic bulletin boards. There was some concern that even when a scientist notes that something is missing, it is not clear what he should do to ensure its inclusion. This situation should be rectified. We agreed to announce our intention to identify old, missed sequences inthe literature. In addition, as part of the collaboration, GenBank will begin a more directed effort to identify missing data.

13. International Advisors

Some discussion time was dedicated to the International Advisory Committee. For future meetings it was felt that:

- The provisionof some intermediate information (such as this report) might be helpful.
- Updates on the main issues raised by the last Advisors meeting would be an obvious theme for the nextmeeting.
- The role of the existing databases with respect to large genome scale sequencing needs to be clarified.
- Some discussion of the public domain status of the databases seems appropriate: should we accept data where the submitter wishes to impose limitations on their availability.

14. Next Meeting

The DNA Database of Japan offered to host the next major collaborative meeting. It will be in Mishima, have a duration of about 1 week, and will take place around 12-16 June 1989.

EMBNet: Network for Molecular Biology in Europe

Chris Sander

BIOcomputing Programme
EMBL, D-6900 Heidelberg, Germany (FRG)

The EMBNet Project

EMBNet is a project to develop the European infrastructure for academic and commercial information services in biotechnology. The project includes the formation of a computer network for the exchange and development of data of importance to molecular biology and biotechnology. The network will consist of a central node at EMBL and regional nodes in the European countries, appropriately staffed and equipped with computing facilities. The network will prepare the infrastructure for the diverse biological and biotechnological information services required by European academic research institutions and the chemical and pharmaceutical industry in the future. During the development, novel aspects of information technology, such as networking, relational database development, computer conferencing and storage media will be applied on a practical and widening scale.

Activity in biotechnology and bioinformatics is intensifying world-wide, thus development of network-based information services is viewed as urgent by public and commercial sectors alike. Maintenance of the basic public databases and related academic and educational activities will eventually be carried out by a projected European Institute of Bioinformatics, while development of bioinformatics and biotechnology products and value-added services will migrate to the commercial sector.

The first trial phase of the network has been initiated and will be operational before the end of 1988.

The need for biotechnology information services

The recent explosion in data as a result of research in molecular biology will have a major impact on research and development in the chemical, pharmaceutical, biotechnological, medical and agricultural technology. As biotechnology based products are developed, computerized access to the underlying databanks becomes increasingly important. Current important databanks include information on: gene and genomic sequences of proteins and nucleic acids, structures of biological macromolecules, human genome maps, genetic diseases, microbial strains, hybridoma, restriction enzymes, cloning vectors, industrial enzymes, taxonomic classification, toxicological data, abstracts from biomedical journals and other areas. DNA sequence data is maintained at present by a worldwide collaboration including DDBJ in Japan, GenBank in the US and the EMBL Data Library in Europe. Protein sequence data are similarly coordinated.

As biotechnological research intensifies and progresses to product development, the commercial sector of information services will develop in parallel. Electronic publishers, database hosts, computer companies, software houses and the chemical, pharmaceutical and agricultural industry will have a major role in the development of the information services market.

Structure of EMBNet

The European Molecular Biology Data Network will consist of EMBL as the central coordinating node connected to a series of national nodes in European countries. The network will be accessible to commercial as well as academic users. It will be based on international communication standards such as ISDN (Integrated Services Digital Network) and OSI (open system interconnect).

Some of the outstanding advantages of a decentralized data network are:

(1) A large pool of expertise can be brought to bear on data flow, research and communication problems in a coordinated yet diversified fashion.

(2) The user community varies from one country to the next, due to language and cultural differences, differences in national computer networks and local facilities. National and regional centers linked to

a central node can specifically address these issues. Special attention can be focussed on the technologically less developed regions.

(3) Remote login across international boundaries is costly and difficult. A system which enables users to retrieve and submit data from a computer within the same country is simpler and cheaper.

EMBNet Network Functions

Activities at central EMBNet node

1. Data Maintenance

Collect and/or maintain sequence data, in continuing close collaboration with GenBank (USA) and DDBJ (Japan). EMBL will make available other data collections important to molecular biology such as genetic and physical mapping data.

2. Data Distribution

Distribute complete releases of all data several times each year on mass storage media such as magnetic tape or CD-ROM. Subsets of data, especially data newly-processed at EMBL between full releases, will be made available via an electronic mail fileserver, and also distributed daily from EMBL to regional nodes

3. Computer Conferencing System

The development of an advanced bulletin-board and conferencing system on the network will provide the vehicle for information exchange between all network partners. Such as system constitutes a prototype for an educational/academic communication system to be established by the European Institute of Bioinformatics.

4. Training

Run training courses for staff of the regional nodes.

5. Database Development

Carry out research in database design for future generations of molecular biology databases, including use of artificial intelligence. Carry out research to design a model computing environment for user access to the databases.

6. Biocomputing

The EMBL Data Library is part of the Biocomputing Programme at the EMBL. Groups within the programme conduct research in several other areas of computational and theoretical biology.

Activities at EMBNet regional nodes

1. Database Access

Make available latest releases of the molecular biology databases and retrieval software to their users.

2. Software Access

Make available licensed and unlicensed molecular biology software for sequence analysis, including locally-developed software. Users should be able to able to copy some software to their local computer as well as to use some of the software packages via remote login to the regional node computer. In this way computing resources are distributed, with remote login work being limited to specialised or computer-intensive tasks (eg database homology searches).

3. Data Entry

Make available EMBL/GenBank Data Submission software. This crucial piece of software is designed to aid sequence and data authors in producing a complete, properly-formatted and internally consistent database entry for automatic transmission, via electronic mail, to the data library. The regional nodes will assist their users in setting up and using the required network communications.

4. User Support

Provide on-site, electronic mail and telephone user support, and organise training courses in the use of molecular biology databases and software.

5. Database Development

Where possible, perform research in aspects of database development or theoretical biology in order to be actively involved in novel aspects of bioinformatics. The results of such research can be shared by other regional nodes (eg. the development of specialised databases related to the main databases).

Outlook

The first four academic nodes are CITI2 in Paris (France), CAOS/CAMM in Nijmegen (Netherlands) and SEQNET in Daresbury (UK). Daily data traffic is to start before the end of 1989.

Nodes in most European countries (members of EMBL and members of the EC) will be added over a period of two years. Industrial links will be developed.

Eventually (about 1993) a proposed European Institute of Bioinformatics (EIB) would be founded by the EC as a European database center for molecular biology and biotechnology and as a training center at various levels. The center would absorb and continue the service functions of separate databases, such as the EMBL Data Library or the Microbial Strain Data Network, and integrate these into a coherent database system accessible throughout Europe via computer links.

The main biological databases, of extreme importance to medicine, technology and the further evolution of life, should always remain in the public domain, accessible to academic, commercial and educational sectors alike and available across all national boundaries.

This document was written in September 1988 for presentation at BIO FAIR TOKYO '88 by Chris Sander with input from the staff of the EMBL Data Library.

ＤＤＢＪ利用初心者講習会開催される

ＤＤＢＪ

第一回講習会を次ページの内容で開催しました。受け入れ可能な人数の２倍以上の申し込みがあり、お断りする人がでました。まことに申し訳ありません。次回の講習会の折に優先致しますのでお許し下さい。

講習会の印象を、参加者の正木先生が蛋白質核酸酵素（ Vol.33 No.13 '88 ）に書いて下さいました。正木先生、及び編集部の許可を得て転載しましたので御一読下さい。

参加者は以下の方々でした。（50音順、敬称略）

氏　名	所　属	講　習	実　習
安孫子宣光	日本大学　松戸歯学部 生化学教室	*	*
伊藤　渉	藤田学園　医学部　総医研 免疫	*	*
江口　幸典	琉球大学　医学部 生化学第２	*	*
尾崎　浩一	大阪大学　理学部	*	
加藤　武司	大阪大学　医学部 放射線基礎医学	*	
木村　晃之	国立がんセンター研究所 生物物理部	*	*
小宮　弘之	名古屋女子文化短期大学	*	*
佐藤　建三	川崎医科大学 生化学	*	
佐藤　勉	広島大学　生物生産学部 微生物生化学	*	*
杉崎　祐司	キッコーマン（株） 第３研究室	*	
高尾　雅	東北大学　抗酸菌病研究所 薬理	*	*
谷　昭義	武田薬品工業(株)　中央研究所 生物工学研究所	*	*
堤　伸浩	東京大学　農学部 放射線遺伝学教室	*	*
徳永　史生	東北大学　理学部 物理	*	
虎沢　慶太	名古屋大学 遺伝子実験施設	*	
堀居　敏彦	塩野義製薬（株）研究所	*	
正木　茂夫	愛知県心身障害者コロニー 発達障害研究所　生化学部	*	
松尾　光一	慶応義塾大学　医学部 微生物学教室	*	*
山西　清文	京都府立医科大学 皮膚科学	*	
横田　匡美	山之内製薬(株)　中央研究所 分子生物学研究部	*	*

国立遺伝学研究所ＤＤＢＪ利用初心者講習会スケジュール

主催：　国立遺伝学研究所　　TEL:0559-75-0771
　　遺伝情報研究センター　DDBJ　　ext.647
会場：　遺伝情報研究センター　４Ｆ
　　遺伝情報分析研究室

6月17日（金）

9:50 - 10:00	挨拶　　遺伝情報研究センター長　瀬野	
10:00 - 10:30	パーソナルコンピュータ，モデムを用いた公衆電話回線利用による計算機アクセスの方法	宮澤、林田
10:30 - 11:00	DDBJnewsアカントの利用法	宮澤、林田
11:00 - 11:30	UNIXシステムの概要（スクリーンエディターviとファイル構成）	宮澤、林田
11:30 - 12:00	ファイル転送について	宮澤、林田

昼休み

13:00 - 13:30	電子郵便を用いた DNAデータのデータバンク(DDBJ, EMBL,GenBank) へのサブミッションの方法	宮澤、林田
13:30 - 15:00	DNA データベースの利用について	宮澤、林田
	UNIXシステム	宮澤、林田
	flat　の利用法	宮澤、林田
	fast ホモロジーサーチ の利用法	宮澤、林田
	qanalys　の利用法	林田、宮澤

休　憩

15:30 - 17:00	VMS システム	宮澤、林田
	システムの概要（スクリーンエディターEDT とファイル構成）	宮澤、林田
	UWGCG　の利用法	藤田*、宮澤、林田
	IDEAS　の利用法	藤田*、宮澤、林田
	PSQ/NAQ　の利用法	宮澤、林田

6月18日（土）

10:00	実習（自由参加）	宮澤、林田

*)所属：分子遺伝研究部門

REPORT

DNA Data Bank of Japan (DDBJ) 利用初心者講習会印象記

正木茂夫

静岡県三島市にある国立遺伝学研究所遺伝情報研究センターより，DDBJ (DNA Data Bank of Japan) の，外部「利用申請書」提出者を対象にした利用初心者講習会を，6月17日(金)と18日(土)の両日にわたって行なうとの連絡があった。筆者はすでに遺伝情報研究センターのコンピュータ「利用申請書」も提出して，ユーザ登録も済ませてあるが，NTT 公衆回線を経由して交信してもどうもコンピュータは相手にしてくれていないらしく，うまく交信できたためしがない，いわば「落ちこぼれ」派。これ幸いと，さっそく手続きをして参加させていただくことになった。

遺伝情報研究センターは，国立遺伝学研究所本館の北側に位置し，1987年1月に竣工したタイル張りの新しい建物である。講習会場としてあてられた4階の遺伝情報分析研究室と称する端末室は，一辺が6メートルぐらいの正方形の明るい部屋で，5台のコンピュータ端末が置かれていた。この講習会への申込み者は，定員枠の20名をはるかに越え50名もに達したそうで，DDBJオンラインサービスに対する期待の大きさがうかがえた。しかし，物理的な制約が大きく，結局20名の定員枠を若干名広げての開催となった。また，最も遠い参加者は琉球大学からであった。

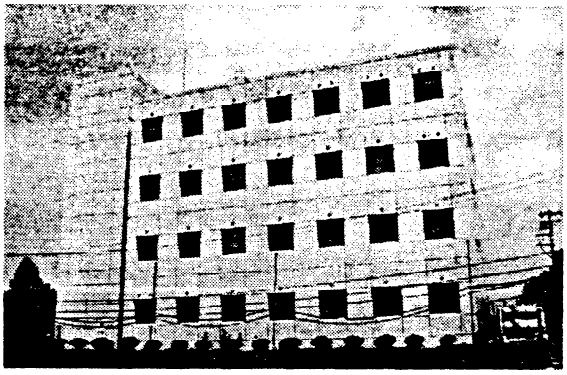


写真 1. 国立遺伝学研究所遺伝情報研究センターの全景

1. DDBJ について

遺伝情報研究センターに設置されている DDBJ については，本誌上[1]にすでに詳しいので，そちらを読んでいただくことにしたいが，一言でいうならば米国 GenBank や欧州 EMBL と肩を並べる，わが国自前の DNA データバンクである。

日ごろわれわれが使用しているコンピュータ可読 DNA 塩基配列データは，テープやフロッピーディスク，または CD-ROM など，さまざまな媒体を経て自由に配布することが許されているが，それらはすべて GenBank と EMBL の手によって入力・データベース化されてきたものであって，そのデータ蓄積に関してはわが国の関与は最近までほとんどなかったといってよい。そこで，わが国としてもデータ蓄積作業の一端をにない，国際的な責任の一部を果たしていくと同時に，DDBJ や他の DNA データベースの国内での配布に責任をもつ機関として，DDBJ が設置されたと聞いている[1,2]。

最近，ヒト全 DNA 塩基配列を決定するいくつかのプロジェクトの具体化にも呼応して，DDBJ，GenBank と EMBL の三者の間で連絡会議がもたれ，DNA データ収集作業の国際間の分業とデータベースの共同構築についての手順が取り決められたことが報告されているが[3]，これをみても DDBJ はすでに GenBank，EMBL と並ぶ位置を占めており，寄せられる国際的な期待の大きいことを示している。しかし，DDBJ には現在までのところ34万塩基程度の DNA 塩基配列データの蓄積しかなく，これは DDBJ が他の DNA データバンクに比

Shigeo Masaki, 愛知県心身障害者コロニー発達障害研究所生化学部 (〒480-03 愛知県春日井市神屋町 713-8)

写真 2. 講習会風景

べて事業規模が小さいうえに，データ収集の開始時期が約5年遅れたのが原因であるが，すでに約1千万塩基を蓄積している GenBank や EMBL とは大きな隔たりがある[1]。

しかし，たとえば今年1年間で世界中の研究者が解析する遺伝子の数は数千個で，塩基数にすると数百万にのぼると推定されており，その結果，蓄積する DNA 塩基配列データは，各 DNA データバンクとも今年1年で前年度比 10 パーセント以上の増大が見込まれている。そして，解析される遺伝子の数は毎年増加の一途をたどっており，ヒト全 DNA 塩基配列決定のプロジェクトの開始も目前である。こういったことを考えれば，今後取り扱うデータ量の増加に応じて DDBJ の事業規模の拡大などが適正に行なわれれば，海外の他の DNA データバンクとの間にある蓄積データ量の差も急速に縮まり，その点でも DDBJ が海外の DNA データバンクに追いつくのは遠くないと思われる。

2. 遠隔地からの利用法

研究者が日ごろ取り扱っている限られた遺伝子のコンピュータ処理については，その DNA 塩基配列をコンピュータ可読の形で入手すれば，かなりのことがパーソナルコンピュータ上でも行なえるようになっている。筆者自身も，いくつかの眼球形成に関与する遺伝子の塩基配列を，友人を介して GenBank より入手し，その制限酵素切断部位を，市販のワードプロセッサの検索機能を用いて検索したり，最近多数出回っている MS-DOS の文字列データを検索したり加工したりするようなプログラムを流用して，簡単なホモロジー検索を行ない重宝している。

しかし研究が進むにつれ，不特定の遺伝子を対象としたホモロジー検索などを行なう必要性が出てくると，筆者のこういった姑息な「裏わざ」もまったく無力となり，しだいに DNA 分析専用のソフトウェアがほしいと思うようになってきた。NEC PC-9801 シリーズには，すでに専用の DNA 分析用ソフトウェアパッケージが数社より市販され，いずれも CD-ROM ドライブに入った DNA 塩基配列データが読めるように作られているので，GenBank など全 DNA データに対してのホモロジー検索までがパーソナルコンピュータで可能になっている。

でも，こういったソフトウェアパッケージは，手軽に使えるのが最大の長所であり，個人使用やごく少人数の研究グループの使用こそ最もふさわしいと思われるが，実際には個人的に購入するには少し高価であり，筆者にはちょっと手が出なかった。そういったときに国立遺伝学研究所遺伝情報研究センターによる DDBJ オンラインサービスが開始されることを知った[1]。

研究者のパーソナルコンピュータで，遺伝情報研究センターのコンピュータを利用するときの最大の課題は，直通の NTT 公衆回線を使えるかどうかである。構内内線電話からでも使えないことはないが，せっかく交信しても「データ化け」が起こっていたり，利用の途中で回線が切れて遺伝情報研究センターに迷惑がかかることがあるので，それなりの決意がいる。

手順としては，まず遺伝情報研究センターより「利用手引書[4]」の送付を受ける。次に「パソコン通信用モデム」を購入する。さまざまな規格があり，DDBJ では，「2400 ボー，MNP 対応」を勧めている。しかし，筆者としては，初心者が個人的に購入するのなら何よりも安価という点で「1200 ボー，AT モデム」とよばれるタイプを勧める。でも，その代わり電話代は少しかさむ(エプソン，タムラ，アイワ，オムロン，NEC 製などがある。最近は2万円以下で買える)。ほとんどの場合，ケーブルも付属しているので，説明書に図示されているとおり結線し，NTT への届けも済ませ，遺伝情報研究センターへ「コンピュータ利用申請書」を送付すれば準備は完了。あとは登録完了の知らせとユーザネームの交付を待つ。

通信ソフトウェアは，DDBJ から NEC PC-9801 シリーズで動くものが配布してもらえるが，高機能すぎてまったく初めての人が使うには骨が折れる。最初だけは簡単に使いはじめられる市販の通信ソフトウェアを使用したほうがよい。でも，遺伝情報研究センターのコンピュータ端末として使える市販品は意外に少なく，筆者と

しては「CCT 98-Ⅱ ((株)技術評論社，定価1万5千円)」を勧める。そして一度は，いわゆる「BBS (Bulletin Boards Service)」とよばれるパソコン通信センターの，いちばん電話代の安いところを捜してアクセスしてみるとよい(「BBS」の電話番号は CCT98-Ⅱの説明書にもたくさん載っている)。どこの「BBS」でもファイルの取り扱いは，遺伝情報研究センターのシステムである UNIX に似せて造ってあるので，ずいぶんファイルアクセスの練習になる。でも，あまりそればかりに夢中になってもいけないが。

3. 講習会のスケジュール

講習会は6月 17 日（金曜）午前9時 50 分より，瀬野悍二遺伝情報研究センター長の挨拶で始められ，すぐに午前中の講義が始められた。まず宮澤三造・林田秀宜両氏の指導でパーソナルコンピュータと NTT 公衆回線使用による遺伝情報研究センターのコンピュータ利用の方法について，具体的な機器の接続方法と操作の方法のデモンストレーションがあり，続いて DDBJ の利用に際して最低限必要な UNIX オペレーティングシステムのファイル構造についての説明と，UNIX 上のスクリーンエディタ“vi”の使用法，さらに UNIX システムでの「電子郵便」の送受信法についての講義があった。話された要点としては，UNIX は極力キー入力や出力を抑えた「寡黙で不愛想な」オペレーティングシステムであるために，MS-DOS などに比べて，とっ付きにくい印象を与えているが，すでにその樹木型ファイル構造は MS-DOS にも取り入れられているので，パーソナルコンピュータを使用している人にもけっして難解なものではないこと，また，どんな状況でもオンラインマニュアルが完備されているので，操作法がわからなくなったときは

```
(1,200ボーのATモデムを使用した交信の例で，PC-9801で通信ソフトウエアを起動した状態から
表示する．下線部はPC-9801より入力する文字を示し，↵はリターンキーを押すことを示す)

:ATPD0559-75-6036↵                    ……(遺伝研遺伝情報研究センターへ電話をする)
OK                                    ……(電話回線が接続された)
[BREAK]↵                              ……(ブレークキーを押してブレーク信号を送る)
:CONNECT 1200                         ……(コンピュータとの交信が1,200ボーで成立した)

niguts
        Welcome to the NIG FACOM-M380Q/UTS V10L32 (system V release 2.0)

(1200 baud) login: smasaki↵           ……(ユーザーネームを入力する)
Password: ddbj007↵                    ……(パスワードを入力する．これは架空のもの)
Terminal type (pc98msdos): ↵          ……(pc-9801ではリターンだけでよい)
                        この間省略
niguts% flat↵                         ……(flatの命令が使えるようにする)
flat% egrep "^ECO" $GENBANK/bacteria.dir |
                                  | egrep -i "heat" | egrep -i "shock" > e.coli-heat-sh↵
      ……(GenBankのbacteriaのディレクトリから，ECOで始まるエントリーを取り出し，
                                  definition行に，heatとshockを含む行をファイルにする)
flat% cat e.coli-heat-sh↵             ……(そのファイルを画面に表示する)

ECOC625    M10419 ds-DNA     96 E.coli heat-inducible promoter for operon encoding he
ECODNAJ    M12565 ds-DNA   1358 E. coli K12 dnaJ gene encoding a heat shock protein,
ECODNAK    K01298    DNA   1917 E.coli dnaK gene encoding the heat shock 70 protein.
ECOHTPR    K02177 ds-DNA   1060 E. coli htpR (heat shock regulatory) gene, complete c
ECOHTPRR   K02178 ds-DNA   1312 E.coli htpR gene coding for heat shock regulatory pro

flat% getgb bacteria.seq ecohtpr >ecohtpr.↵
      ……(GenBankのbacteriaの配列データからエントリーECOHTPRを取り出しファイルにする)
flat% pg ecohtpr.↵                    ……(そのファイルを画面に表示する)

LOCUS       ECOHTPR      1060 bp ds-DNA            BCT       01-JUL-1985
DEFINITION  E. coli htpR (heat shock regulatory) gene, complete cds.
ACCESSION   K02177
KEYWORDS    heat shock protein.
                        この間省略
                        この間省略
     961 aattgcgtgc tgccattgaa gcgtaatttc cgctattaag cagagaaccc tggatgagag
    1021 tccgggggttt ttgtttttttg ggcctctgta ataatcaatt
//

flat% fromgb ecohtpr. | rsites $FLAT/lib/enzymes/avail.enz >ecohtpr.rsites↵
    ……(そのファイルをスタンフォードフォーマットに変換し，制限酵素切断部位を検索する)
flat% rmap < ecohtpr.rsites↵          ……(その結果を画面にグラフで表示する)

* ECOHTPR      1060 bp ds-DNA            BCT       01-JUL 0 1060
                      168       345       522       698       875       1052
        AccII   5 --------+-1-------+---------+---------+1---1--1+---1-----+
        AflII   1 ---1-----+---------+---------+---------+--------+---------+
                        この間省略
                        この間省略
       EcoRII   5 ---------1---------+---------1---------+--1-----+---------1-+---------1-+
        EcoRV   1 -------1--+---------+---------+---------+--------+---------+---------+
       EcoT14   1 ---------+----1----+---------+---------+--------+---------+---------+
         EspI   1 ---------+---------+---------+---------+--------+---------+--1------+
                        この間省略
                        この間省略
        SstII            StuI          Tth111I           XbaI            XhoI
         XmaI           XmaIII           XorII
//

flat% grep -i ecorv ecohtpr.rsites↵   ……(EcoRVについての結果のみを表示する)

  EcoRV GATATC  110-115

flat% exit↵                           ……(flatを終了する)
niguts% logout↵                       ……(交信を終了する)

smasaki logout, Mon Jul  4 14:15:34 JST 1988

NO CARRIER                            ……(電話回線が切断された)
:
```

〔例1〕

とにかくヘルプ命令を呼べば，いつでもオンラインで説明が受けられることなどが話された。

午前中の講義で，筆者にとっても今後とりわけ重要になると思われたことは，国際 UNIX コンピュータネットワークを用いた「電子郵便」の送受信法であった。最近ではいくつかの雑誌（*Nucleic Acids Research* 誌など）で，DNA 塩基配列を含む論文を投稿する際には，前もってその配列のデータをコンピュータ可読の形（具体的にはフロッピーディスク）でその雑誌を担当している DNA データバンク（*Nucleic Acids Research* 誌の場合は EMBL）に送付して，その受領書を添えて論文を投稿することになっている。それを航空便で行なおうとすれば最低1カ月は必要であるが，「電子郵便」で DDBJ を経由して EMBL へ送付すれば遅くとも2日後には DDBJ で受領書を手にすることができ，そういったときには申し出れば，DDBJ としても協力を惜しまないとのことであった。

午後からは，遺伝情報研究センターのコンピュータで使うことのできる DDBJ の「DNA 塩基配列処理プログラム群」として“flat”と“qanalys”が宮澤・林田両氏から，また，“UWGCG”と“IDEAS”および“NAQ”が分子遺伝研究部門の藤田信之氏から紹介され，それらの簡単な使用法の説明があった。これらのプログラム群の相互の関係を述べることは筆者にはいささか困難であるが，大略すると以下のようになると思う。すなわち，これらのうちで“flat”と“qanalys”はUNIX システム上で動き，2つで1つのプログラムともいうべき関係にある。“flat”はおもに DNA データバンクより任意の塩基配列を取り出すために使用し，取り出した塩基配列の解析は，

```
（設定は同じ）

:ATPD0559-75-6036
OK
[BREAK]
:CONNECT 1200


niguts
        Welcome to the NIG FACOM-M380Q/UTS V10L32 (system V release 2.0)

(1200 baud) login: smasaki
Password: ddbj007
Terminal type (pc98msdos):
                              この 間 省 略
niguts% nigvms                          ……（交信するコンピュータをVAXに変更する）

telnet nigvms
Trying...
Connected to nigvms.
Escape character is '^D'.

       Welcome to the NIG MicroVAX-II/VMS V4.4

Username: SMASAKI                       ……（再度ユーザーネームを入力する）
    Last interactive login on Wednesday, 29-JUN-1988 09:31
Terminal type (set term/inq) ? vt100            ……（端末はvt100と入力する）
 >
Use "getinfo" at the niguts to get information; type "telnet niguts" to login.
                              この 間 省 略
nigvms$ naq genbank             ……（NAQの命令が使えるようにし，GenBankを呼び出す）

Nucleic Acid Query System V3/2.4

NAQ> find e. coli heat shock
  ……（GenBankの中から，definition行にE., coli, heat, shockを含むエントリーを検索する）

ECOC625     E.coli heat-inducible promoter for operon encoding heat shock protein
ECODNAK     E.coli dnaK gene encoding the heat shock 70 protein
ECOHTPR     E. coli htpR (heat shock regulatory) gene, complete cds
ECOHTPRR    E.coli htpR gene coding for heat shock regulatory protein F33.4 and
M12565      E. coli K12 dnaJ+ gene encoding a heat shock protein
  5 entries found

NAQ> quit                               ……（NAQを終了する）
nigvms$ ideas seqman                    ……（IDEAS中のSEQMANの命令が使えるようにする）
% get genbank ecohtpr >ecohtpr.
                        ……（GenBankの配列データからエントリーECOHTPRをファイルにする）
% type ecohtpr.                         ……（そのファイルを画面に表示する）

LOCUS       ECOHTPR      1060 bp ds-DNA             entered   07/01/85
DEFINITION  E. coli htpR (heat shock regulatory) gene, complete cds.
                              この 間 省 略
     1021 TCCGGGGTTT TTGTTTTTTG GGCCTCTGTA ATAATCAATT
//

% end                                   ……（SEQMANとIDEASを終了する）
nigvms$ uwgcg                           ……（UWGCGの命令が使用できるようにする）

                        Welcome to the
              University of Wisconsin Genetics Computer Group
                        Version 3.0, June 1985
```

〔例2〕 その1

```
nigvms$ fromgenbank↵          ……(GenBankからUWGCGフォーマットへの変換処理を行なう)

 Reformat what GenBank data file?  ecohtpr.↵          ……(ファイル名を入力する)
       ecohtpr.  1060 bp.
 reformatted: ecohtpr.
 total files: 1
 total bases: 1060

nigvms$ map↵                  ……(制限酵素の切断部位の検索を行なう)

MAP is a sequence display tool that shows the sequence and its complement
with the restriction map above and possible protein translations below.
Local data files:  Enzyme.dat (enzyme names and recognition sites)
Command Line Switches
        /SIX            shows six-base cutters when enzyme name is "*"
        /ONCe           shows only enzymes that cut once in the chosen range
        /LINear         treats the sequence as linear (default)
        /CIRcular       treats the sequence as circular
        /WIDth          allows choice of number of characters per line
        /MISmatch       allows mismatches in site search
        /APPend         appends the enzyme data file to MAP's output
 (Linear) MAP of what sequence ?  ecohtpr.↵    ……(ファイル名を入力する)
                Begin (* 1 *) ?  1↵           ……(データの開始塩基番号を入力する)
              End (*  1060 *) ?  1060↵        ……(データの最終塩基番号を入力する)
 Enter enzymes:  Use a "*" to get all, a "$" to get none
 or enter the names individually, one per line, ending
 the list with a blank line.
                  Enzyme(* * *):  ecorv↵     ……(検索する制限酵素名を入力する)
                          Enzyme:  ↵         ……(終了はリターンのみ)
 What protein translations do you want:
     a) frame 1   b) frame 2   c) frame 3
     d) frame 4   e) frame 5   f) frame 6
     t)hree forward frames   s)ix frames   o)pen frames only
     n)o protein translation   q)uit
 Please select (capitalize for 3-letter) (* t *):  ↵      ……(通常はリターンのみ)
 What should I call the output file (* ecohtpr.map *) ?  TERM↵
                                                  ……(画面表示の時はTERMとする)
 MAP of: ecohtpr.  check: 6042  from: 1  to: 1060
locus       ecohtpr      1060 bp ds-dna              entered   07/01/85
definition  e. coli htpr (heat shock regulatory) gene, complete cds.
accession   k02177
keywords    heat shock protein.
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
                                                          E
                                                          C
                                                          O
                                                          R
                                                          V
        GTTGCTCTTAAGCTCTGGCACAGTTGTTGCTACCACTGAAGCGCCAGAAGATATCGATTG
     61 ---------+---------+---------+---------+---------+---------+ 120
        CAACGAGAATTCGAGACCGTGTCAACAACGATGGTGACTTCGCGGTCTTCTATAGCTAAC
a       V  A  L  K  L  W  H  S  C  C  Y  H  *  S  A  R  R  Y  R  L
b        L  L  L  S  S  G  T  V  V  A  T  T  E  A  P  E  D  I  D  *
c         C  S  *  A  L  A  Q  L  L  L  P  L  K  R  Q  K  I  S  I  E
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
b       P  G  F  L  F  F  G  P  L  *  *  S  I
c        R  G  F  C  F  L  G  L  C  N  N  Q  ?

nigvms$ LOG↵                  ……(UWGCGを終了し，VAXとの交信を終わる)

  SMASAKI      logged out at  4-JUL-1988 14:53:49.50
Connection closed by foreign host.

niguts% logout↵

NO CARRIER
:
```

〔例2〕 その2

“qanalys” で行なうように作られている。UNIX 上で動くことから，きわめて高速で処理ができるのが特徴であるが，現在開発中の部分もあり，まだできない処理もいくつかあるようである，他方，“UWGCG” と “IDEAS”，“NAQ” は，VAX コンピュータ上で走る世界中で最も有名なプログラム群である。“UWGCG” は DNA データバンクのデータ操作機能をもち合わせていないが，そのほかはともに「DNA 塩基配列処理プログラム」としてほぼ完璧な機能をもっている。しかし，処理速度はそれほど速くない。最後に，いくつかの例題について簡単な実習を行なった。実習の事例（交信例）を示す。〔例 1〕は UNIX システムの “flat”，〔例 2〕は UNIX を経由して VAX 上で “NAQ”，“IDEAS” と “UWGCG” を用いて，ともに大腸菌熱ショック蛋白質 DNA 配列データと制限酵素切断部位の検索を行なってみたものであるが，その雰囲気のごく一部はつかんでいただけると思う。

それぞれのプログラム群で使用できる命令と処理の詳細については，誌面に限りもあり，ここで紹介することはできないので，一度試みてみようという方はぜひ遺伝情報研究センターに問い合わせてみていただきたい。英文の詳細な「プログラム使用説明書」が送付されるはずである。もっとも，講習会の参加者にも事前にこの「プログラム使用説明書」が送付され，筆者も目は通したが，ちんぷんかんぷんで結局はほとんど理解できなかった。でも，この講習会に参加して実際に自分の手で端末を操作してみると，それなりに使用できるような気分になれて幸せであった。実践は偉大である。

4. オンラインサービスに望むこと

昨年，遺伝情報研究センターがコンピュータのオンラインサービスを開始

するまでは，DDBJ の DNA データにアクセスするのは DNA 塩基配列処理プログラムの使用法に習熟した研究者が大部分であり，しかも直接，端末の使用か遺伝研講内の内線回線を通じての使用がほとんどであったはずである。それが最近は内部の使用者に加わって，筆者のようなよくわかっていない初心者までが，NTT からの目の玉の飛び出るような請求書におびえつつも，遠隔地からよろよろとアクセスするようになった。そして，そういった初心者ほど，自分がうまく操作できないことの原因をいつもコンピュータシステムのせいにし，ホストシステムに過大の要求をして，なにも恥じることがないから始末が悪い。そして筆者もその例に漏れず，講習会のあと開かれた懇親会の席上で，臆面もなくいろいろな要望をしてしまった。

まず1つめは，“flat” や “qanalys” などは多くの引数をコマンドラインで入力するために，オンラインマニュアルが用意されてはいるものの，慣れないと使いにくく，それが UNIX システムの仕様なのだと聞かされても，初心者には使いづらい。その点では VAX 上で走る “UWGCG” や “IDEAS” のほうが，速度は遅くても対話方式で操作できるので使いやすい。UNIX でも，初心者向けに対話方式で操作ができるようにしてほしい。2つめは，遠くから，アクセスしているときには英語ではなくて視認性のよい日本語のオンラインヘルプを表示してほしい。3つめには，それぞれの DNA データバンクのデータフォーマットが細部で異なっているのは，各々よさもあり歴史もあるから仕方のないことだろうが，処理ごとにフォーマット変換を要求されるのは初心者にはつらい。さらには，最寄りの「国立大学計算機センター」から DDBJ へアクセスしたいということや，パーソナルコンピュータ向けの DNA データ処理プログラムの配布サービスもお願いしたいなどと，長々と申し上げた。

瀬野センター長と宮澤氏は，こういった筆者の身勝手な要望を黙って聞いておられたが，最後には1つ1つ返答してくださった。それらを以下に要約する。

（1） UNIX 上の “flat” や “qanalys” もコマンドライン以外に，対話方式でも使用できるような形に改良して，初心者でも使いやすいように今後改善したい。

（2） DDBJ, GenBank, EMBL 各々の DNA データバンクのデータフォーマットは，見直しが急がれており[3]，近く統一され，フォーマット変換は不要になる。

（3） ごく最近「国立大学計算機センター」の専用回線（N1-NET）を経由して DDBJ が使用できるようになったので，希望者は連絡してほしい。

（4） パーソナルコンピュータの処理速度が向上しており，DNA 塩基配列データ処理をその上で行なう傾向はますます強まるものと思われるので，パーソナルコンピュータ向けプログラムの配布サービスも考えている。たとえば利用者の開発した DNA データ処理に有用なプログラムを寄せていただき，DDBJ を通じて配布するなどしたい。

いちばん最後に，DDBJ を使いやすいものにし，GenBank や EMBL に負けない DNA データバンクにしたいと考えているので，皆さんにどんどん利用していただき，要望や意見を寄せていただくよう一層のご協力をお願いする，と述べられた。

利用者講習会と懇親会の終わったころには，初夏の長かった日も落ちて帰りのバスはすでになく，遺伝研の長い暗い坂道を市街地に向けて歩いた。きつい日程の1日であったが，たいへん実りの大きい1日でもあった。

6月 18 日(土)は，1日がかりで実習が行われた。

最後にこの「DDBJ 利用初心者講習会」を企画し，2日間を充実したものにしていただいた国立遺伝学研究所の瀬野悍二遺伝情報研究センター長，遺伝情報研究センターの宮澤三造・林田秀宣両氏と職員の方々，ならびに分子遺伝研究部門の藤田信之氏に心から感謝する。

文 献

1) 丸山毅夫：本誌，**33**, 268-270 (1988)
2) 内田久雄：蛋白質・DNA のデータバンクと情報解析（本誌別冊 No. 29），pp. 159-162 (1986)
3) Soll, D. *et al.*: *Science*, **240**, 375 (1988)
4) 遺伝情報研究センター DDBJ 編：「国立遺伝学研究所 DNA Data Bank of Japan 共同利用電子計算機利用の手引」

DDBJ 計算機において利用可能なデータベース

Versions of Databases

DNA data base	Release date		
DDBJ	4	01/89	
EMBL	18	02/89	
GENBANK	59	03/89	
GENBANK	52	08/87	for msdos floppies
HIV-N	88.2a	1988	Human Retroviruses and AIDS
KABAT		1983	Seq. of Imm. Interest
NBRF	34	11/88	
Codon usage			
Codon-usage	1		GenBank R.47 is used.
Protein data base	Release date		
DDBJ	4	01/89	translated from DDBJ
HIV-P	88.2a	1988	Human Retroviruses and AIDS
KABAT		1983	Seq. of Imm. Interest
NBRF-PIR	19	12/88	
PGtrans	35	09/85	translated from GenBank
PRF		03/88	Peptide Research Institute
SWISSPROT	7	04/88	
Protein structure data base			
PDB	40	04/87	
LIMB	1	02/88	Listing of Molecular Biology Databases

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

米国からGenBank，NBRF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、希望者に配布した。その他蛋白質データベース（NBRF-PIRデータベース ）も希望者には配布している。配布媒体は GenBankの場合は磁気テープとフロッピーディスク、その他は磁気テープのみである。 GenBankフロッピーディスクに付属して配布されるIBM-PC用の検索プログラムは、NEC-PC9800用に移植しデータと共に配布した。配布形態は定期もしくは一次配布である。磁気テープの配布総数は1116本（1987年 504本、1988年 612本）である。フロッピーディスクの配布枚数は1296枚（1987年 724枚、1988年 572枚、GenBank 24枚／件、DDBJ 2枚／件）である。今年度の配布実績の詳細は以下のようである。

	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）			
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)
	（ ）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（更新者34）	11（更新者6）	75
Kermitプログラム配布数　1987年度	32	5	37
Kermitプログラム配布数　1988年度	11	8	19
VT emulator 配布数　1987年度	11	1	12
VT emulator 配布数　1988年度	7	5	12

GenBank　　(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	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	5	3	0	3	5	6

(57版は、配布途中)

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

EMBL　　(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	86/04	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	1	2	0	1	1	3

(17版は、配布途中)

DDBJ　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
1	0	0	0	0	0	0
2	0	1	3	2	3	3
3	0	1	4	2	4	3

NBRF　　(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
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

NBRF　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
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

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

（18版は、配布途中）

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

（18版は、配布途中）

DNA、蛋白質 データベースの配布について

1. この度、PIR の Directorである Barker博士より今後 JIPID (International Protein Information Database in Japan) が、 PIRデータの日本での"official agency"として機能するとの通知を受けました。 そのため今後 JIPIDが、 PIRデータの日本での２次配布を引き受けることになるであろうと思いますので予めお知らせする次第です。 勿論その際には DDBJ としても配布がスムーズにいくよう JIPID に引継ぎますのでその点はご安心ください。
参考のため以下に JIPID の連絡先を記します。

代表　次田　皓　教授
International Protein Information Database in Japan
〒278　千葉県野田市山崎2641
東京理科大学生命科学研究所
電話：0471-24-1501 内線 5001

2. GenBank は、1989年 6月頃より、CDROM 版をリリースする予定である。

3. PGTran は、版が古くなりすぎたため配布を中止する。

4. EMBLデータベースより翻訳した蛋白質データベース SwissProt を新たに配布する。

5. DDBJ, GenBank, EMBL 配布データに関して以下のような変更があった。

GenBank Release 54 (12/87)
- Short directory files の書式が変更された。（配列の長さが行末に記載）
- 著者名のインデックスの方法が変わった。論文で記載されている著者名を採用

GenBank Release 54 (12/87)、DDBJ Release 4 (1/89)
- Features tableを改変しSites tableはNew features table に吸収された。

GenBank Release 55 (3/88)、 DDBJ Release 4 (1/89)
- LOCUSレコードの書式が変わった。生物グループ名（3 文字表現）を含むこと、また日付の書式がEMBL方式に変わった。
- データ注釈のレベルを表示するSTANDARDレコードがREFERENCEレコードの一部として追加された。

GenBank Release 56 (6/88)
- Sequence data files の先頭にあったShort directoryを廃止した。
- 全てのファイルの一行目の 1-9 カラムにファイル名を表示する行を付加。

GenBank Release 58 (12/88)
- Long directory filesの配布中止。

EMBL Release 15 (5/88)
- Eukaryotic promoter database (EPD) が付加的データベースとして Restriction enzyme databaseと共に配布開始。

EMBL Release 16 (8/88)
- index filesの書式が GenBank方式に類似のものに変更。

6. 今後のDDBJ, GenBank, EMBL 配布データに関して以下のような変更が予定されている。

- 論文雑誌名の省略方法を変更する予定。
  データバンク (DDBJ, GenBank, EMBL, PIR)は National Library of Medicine (USA)と共同で省略名の標準化を進めている。
- EMBL は配列データで U の代わりに T を使用する計画である。
- New features table の採用

7. 配布用磁気テープについてお願い

データ配布のための磁気テープが返却されないケースが特に大学関係者に多く、配布業務が円滑に行われなくなっています。データ量の増加に伴い、配布に必要な磁気テープ数も多くなりました。配布した磁気テープは、できるだけ早くコピーして返却して下さるようお願い致します。

また、定期配布を御希望の方は、予め磁気テープをお送り下さい。磁気テープをお預け頂けない場合は、配布が遅れる場合があることをご承知おき下さい。現在、配布に必要な磁気テープ数は以下の通りです。

	Density: 6250bpi	Density: 1600bpi
GenBank	2,400ft x 1	2,400ft x 3
EMBL	2,400ft x 1	2,400ft x 3
DDBJ	600ft x 1	600ft x 1
SwissProt	1,200ft x 1	2,400ft x 1

なお現在データの配布は No Label及び固定長の形式でのみ受け付けています。VAX/VMS Backupフォーマット, Tar フォーマットでの配布も考慮中ですのでしばらくお待ち願います。

DDBJ/EMBL/GenBank Sequence Data Submission Form 配布のお知らせ

計算機可読な形でDNA データ提出をお願いするためにData Submission Formをフロッピー(MS-DOS 2HD/2DD)で配布しています。Form はDDBJ, EMBL, GenBank共通です。データ提出される方でまだお持ちでない方はソフトウエア申し込み書でご請求下さい。

ソフトウエア配布のお知らせ

これまで DDBJが配布しているソフトウエアは、端末エミュレータ（Kermitの新版、VT emulatorの新版）でしたが、今度これに加えて新しく UNIX（Sun OS/System V/BSD 4.2/BSD 4.3)用のDNA 及び蛋白質データベースのための検索解析プログラムパッケージ FLATを配布することになりました。配布するものは DDBJ計算機の上で稼働しているFLATとほぼ同じです。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。

マニュアル配布のお知らせ

DDBJ計算機で稼動している DNA, 蛋白質解析ソフトウエアのマニュアルを一部作製致しました。未だ十分とは言えませんが、利用の手引きとともに御利用下さい。配布を御希望の方は、巻末のDDBJニュースレター及びマニュアル申し込み書をお送り下さい。

なお、配布可能なマニュアルにつきましては、申し込み書を御参照下さい。

学会デモンストレーション

DDBJのデータ収集について協力を求めるため、以下の学会でデモンストレーションを行いました。

日本癌学会	1988年 9月20日 － 22日
日本生化学会	1988年10月 5日 － 6日
日本分子生物学会	1988年12月20日 － 23日

DDBJ 見学者一覧（1988年 4月 - 1988年 11月）

1988年 6月 1日	東海地区国立大学等附置研究所所長会議出席者	8 名
1988年 6月 8日	文部省学術国際局研究機関課長	1 名
1988年 9月19日	静岡県健康国際フォーラム参加者 大学教授、医師等	30 名
1988年10月19日	Dr. James Cassatt GenBank project Officer National Institute of General Medical Sciences NHI, U.S.A. Seminar: "Nucleic Acid Databases - Vision for the Future"	
1988年10月24日	Dr. Christian Sander Biocomputing Programme EMBL, FRG Seminar: "EMBNet: Network for Molecular Biology in Europe"	
1988年11月 4日	ＪＩＣＡ研修生	8 名
1988年11月21日	東京大学工学部工業化学科・合成化学科職員	21 名

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- 24日	DDBJ-EMBL-GenBank meeting 遺伝研で開催予定。

編集後記

ニュースレターの完成が遅れ申し訳なく思っています。 9割は 1988年 12月のデータバンク運営委員会の時に完成していたのですが、1989年に入り諮問委員会等もあり、伸び伸びになってしまいました。そのため、第一回、第二回国際諮問委員会報告が同居するはめとなりました。

学会でデモンストレーションをしたためでしょうか？ だいぶ DDBJ にデータの提出が増えてきました。 GenBank, EMBL へ提出する場合もどうぞ DDBJ へ提出してください。電子メールで転送いたします。1989年 6 - 7月からは、GenBank 入力担当の場合でも DDBJ に提出されたものは DDBJ が入力することになります。日本で生産されたデータを DDBJ が全て処理できる日も近いと思います。

日本でもヒトゲノム解析の計画が発足しつつあります。ゲノム解析は、データ管理に関して、データ量の増加に対応できるだけでなく質的変化をも要求します。データバンクもこのような新しい要求に応えるべくデータベースの再構築を実施中です。生物情報という性質上、データベースは拡張性にすぐれ、その改変も容易でなければなりません。 又、Fuzzy Information（あいまいな情報）や、遺伝情報の多型（狭義には DNA配列の多型）をいかにデータベース化するか等、多くの解決せねばならない問題が存在します。これらは多分野のデータベースでは見られない遺伝情報に特異な問題であり、現在のデータベース科学でも未解決な問題のようです。その意味でデータベース学としての基礎的研究が必要のようです。専門家の助けが求められます。しかし待っているわけにもいきません。配列データベースとしては適さない面がありますが、現時点では最善と思われる関係データベースへの移行を実施中です。残念ながら DDBJ としてはデータベースを構築するのに必要なソフトウエアを独自に開発するにはスタッフが足りません。GenBank, EMBL で開発されたソフトウエアの移植が手一杯というところです。関係データベース移行後は DDBJ, GenBank, EMBL間でのデータ交換が容易になります。一方、本質的に分散データベースですのでデータ交換のために高速な通信回線、ネットワークが必要となります。 DDBJ は米国の ESNET, Internetへの接続を今年度計画しています。Internet には米国内の大学、研究所、企業の何千もの計算機が接続されていますが、そのいずれの計算機とも login、ファイル転送が可能となります。このようなネットワークはゲノム解析計画でも必須と思われます。

さて、ようやく New feature table が完成しました。使用は容易さから関係データベース移行後ということになります。1989年末には、新しい Feature tableでデータを配布することになるでしょう。

データバンクはもう一つ重要なプロジェクトを実施中しつつあります。それは、研究者自身によるデータ入力を支援するソフトウエアの開発です。GenBank (IntelliGenetics) が開発を担当していますが 4 月にもその version 1 がリリースされます。Beta version を試みましたが、非常に使い易くその使用が待たれます。New feature table を使用しているためこのソフトウエアの使用も関係データベース移行後になります。このソフトウエアは IBM-PC 用ですので、それまでに NEC PC9800 シリーズに移植を計画しています。

以上でお判りのように、関係データベースへの移行が criticalです。しかし 2 名のスタッフには多過ぎるほどの仕事量です。 DDBJ に講習会開催を望む声をよく耳にします。我々もデータ解析のソフトウエアも開発したいのですがなかなか時間がありません。

データ管理は皆さんの目につきにくく目立ちません。しかし GenBank, EMBLデータベースを御使用の方は DDBJ が入力したデータを使用しています。Accession number が D で始まるものは全て DDBJ が入力したものです。御理解をお願いします。

宮澤　三造

遺伝研ＤＤＢＪの利用に関して

1988年12月
国立遺伝学研究所
遺伝情報分析研究室
ＤＤＢＪ係

1986年 4月に発足した DDBJ（DNA Data Bank of Japan）の活動は大きく分けて以下の4 つがあげられます。

1）DNA sequence の収集と入力
1986年12月に入力を開始し、1988年 7月にRelease 3.0 をリリース。1989年 1月にRelease 4.0 をリリース予定。GenBank，EMBLへのデータ送付の中継もします。

2）DNA（GenBank，EMBL，NBRF），Protein（SwissProt)データの配布
1986年 4月より磁気テープで定期的に配布。GenBankデータはフロッピー版も配布。

3）遺伝研共同利用計算機を用いての DNAデータのオンライン利用のサポート
外線電話、DDX-パケット回線が敷設され、また DDBJ 計算機システムの所外利用に関しする利用規定も定まったので、計算機のオンライン使用を1987年 9月より開始している。営利機関、非営利機関の別なく利用可能。

4）ニュースレターの発行等の広報活動
1987年 2月に NO. 6、1987年 11月 No. 7 発行。

計算機を利用する場合は以下の機材、手続きが必要です。（利用の手引参照）

1）端末、もしくはパーソナルコンピューターと端末エミュレータープログラム。
NEC-9800用端末エミュレータープログラム（ＶＴエミュレーター）とデータ転送用端末エミュレータープログラム(Kermit)を希望者には配布しますので、申し込み書をお送りください。

2）全二重式モデム
通信速度の速い（2400 bauds）もので、できればエラー修正をする MNP モデムを推薦します。利用の手引を参考にして下さい。

3）端末とモデムを接続する RS232C ケーブル
パーソナルコンピューターの付属品として購入できます。

4）外線、不可能な場合は内線電話（モデムを接続するため電話線と受話器との接続をモジューラー方式に変換すること。部品を購入すれば個人で簡単にできます。）
遠隔地ではパケット通信を利用するのが割安かも知れません。外線電話を利用する第二種パケット通信回線の申請例は利用の手引を御覧下さい。

5）DDBJ 計算機利用申請書提出
当面試験利用ですので、利用料金は無料です。

営利機関、非営利機関の別なく研究者であれば利用できますのでどうぞご利用下さい。

DDBJ 計算機システムは、UNIXの電子郵便、電子掲示板ネットワーク JUNETに加入しています。DDBJ，GenBank，EMBL へのデータサブミッションに電子郵便をご利用下さい。当然計算機可読ですので、GenBank，EMBLもそれを望んでおります。現在、外国向け電子郵便の発着信は有料となっていますが、GenBank，EMBLへのデータ送付及び連絡は、DDBJが自動的に転送します。詳しくは利用の手引を参照下さい。電子郵便は国内外数時間で到着しますので、非常に便利です。国内、外の研究者との研究連絡用としてもご利用下さい。

DNA データ送付用ならびにDDBJ活動に関する広報のために、特別のアカント(ddbjnews)を設けました。詳しい情報を得るには、計算機にログイン後 "getinfo"コマンドを利用して下さい。又、このコマンドを使用すると、システム及びデータベースに関する情報はもとより、データ配布、ソフトウェア配布申し込み用紙等も手に入ります。申し込み用紙にエディターを用い記入の上、ddbjあて電子郵便で申請書を送り下されば、それに従い処理致します。DDBJ側の労力を減らすためにも、このような事務にも計算機を利用下さるようお願い致します。御意見、質問等もメールの形でお寄せ下されば助かります。計算機へのアクセスの方法は、利用の手引を参照下さい。

計算機導入から日が浅く、UNIXシステムでは DNAデータ解析用プログラムは、今後、順次開発する予定で、現在では十分ではありません。不十分な面は、当面 VAX/VMSを使い、補って頂くようお願い致します。M380Q/UTS とVAX/VMS のほとんどのプログラムに関してオンラインマニュアルが利用できます。今後印刷物としての配布も考えていきたいと思っています。

資料をお望みの方は、ニュースレター配布申し込み書をお送り下さい。

MNPモデム一覧表（平成元年４月現在）（資料を提供下さった東京大学大型計算機センターに感謝いたします。）

モデム名	MNP クラス	V.21	V.22	V.22 bis	NCU AA	NCU MA	NCU MM	Hayes AT	V.25 bis	独自	問い合わせ先		定価（参考）	備考
DATA EXPRESS	3	○	○	○	○	—	—	○	○	—	0587-55-2201	サン電子	128000	
MNP 24	3	○	○	○	○	—	—	○	○	—	03-498-5050	日本ダイレックス営業部	128000	
TrailBlazer T2000PC98	3	—	○	○	○	○	○	○	—	—	03-294-8238	コネクト	148000	18,000bps PC98ｼﾘｰｽﾞ
TrailBlazer T2000PC	3	—	○	○	○	○	○	○	—	—	03-294-8238	コネクト	148000	ﾎﾞｰﾄﾞ J3100,AXｼﾘｰｽﾞ
TrailBlazer T2000SA	3	—	○	○	○	○	○	○	—	—	03-294-8238	コネクト		
MTLOOPER 2400	3	○	○	○	○	○	○	○	○	—	0587-55-2201	サン電子	198000	18,000bps
TrailBlazer T2000	3	—	○	○	○	—	—	○	—	—	03-295-0058	シーエス情報機器	198000	18,000bps
ACCESS 24S	3	○	○	○	○	○	○	○	—	—	03-544-8210	日商エレクトロニクス	198000	
TrailBlazer T2000SA	3	—	○	○	—	—	—	○	—	—	03-341-2566	丸菱エレクトロニクス	198000	18,000bps
ACCESS 24V	3	—	○	○	○	—	—	○	—	—	03-544-8210	日商エレクトロニクス	200000	
2400VP	3	—	○	○	○	—	—	○	—	—	03-498-5050	日本ダイレックス営業部	240000	
PVA24MNP4	4	○	○	○	○	○	○	○	—	—	03-835-1201	アイワ	49800	
MD2400B	4	○	○	○	○	○	○	○	○	—	03-436-7266	立石電機	49800	
COMSTAR 2424AT/4	4	—	—	○	○	—	—	○	○	—	03-798-7846	日本電気テクノマーケティング	50000	
ACER MODEM2424	4	—	○	○	○	○	○	○	—	—	045-433-1211	フジ・ディジタル・システムズ	59800	
DIALNET3000 MODEL3124EH TYPES1	4	—	○	○	○	○	○	○	—	—	03-343-1811	アイ・テイ・エスジャパン	120000	
CTS2424AMH	4	—	○	○	○	—	—	○	—	—	03-366-9741	コンピューターダイナミクス	125000	
CTS2424AM	4	—	○	○	○	—	—	—	—	○	03-366-9741	コンピューターダイナミクス	125000	
CTS2424AM	4	—	○	○	○	—	—	—	—	○	03-814-3081	共和理化	125000	
CTS2424AMH	4	—	—	○	○	—	—	○	—	—	03-814-3081	共和理化	125000	
COMSTAR 2424	4	—	○	○	○	○	○	—	○	—	03-798-7846	日本電気テクノマーケティング	128000	
AX2400J	4	—	○	○	○	○	○	○	—	—	045-433-1211	フジ・ディジタル・システムズ	129000	
AX2400J	4	—	○	○	○	○	○	○	—	—	03-341-2566	丸菱エレクトロニクス	149000	
OSI8224	4	○	○	○	○	—	—	○	○	—	06-443-3260	インテグラン	158000	
VenTel 2400-34	4	—	○	○	○	—	—	○	—	○	03-295-0058	シーエス情報機器	160000	
CDS224Series Ⅱ	4	○	○	○	○	—	—	○	—	○	03-436-9574	兼松エレクトロニクス	170000	
AX2400	4	—	○	○	○	—	—	○	—	○	03-220-0535	ヒューコム	180000	
VenTel 2400-34	4	—	○	○	○	—	—	○	—	○	03-264-2431	三協インターナショナル	180000	
OSI8226	4	—	○	○	○	—	—	○	○	—	06-313-0671	インテグラン［高千穂交易］	180000	
AX2421	4	○	○	○	○	—	—	○	—	○	03-220-0535	ヒューコム	200000	
CDS296 TRELLIS MODEM	4	—	—	—	○	—	—	○	○	○	03-496-9574	兼松エレクトロニクス	398000	
VenTel EC18K-34	4	—	○	○	—	—	—	○	—	—	03-264-2431	三協インターナショナル	740000	18,000bps

モデム名	MNP クラス	V.21	V.22	V.22 bis	NCU AA	NCU MA	NCU MM	Hayes AT	V.25 bis	独自	問い合わせ先		定価（参考）	備考
PVA24MNP5	5	○	○	○	○	○	○	○	—	—	03-835-1201	アイワ	54800	
ACER MODEM2424	5	—	○	○	○	○	○	○	—	—	045-441-8611	コア	59800	
PM2400F	5	—	○	○	—	—	—	○	—	—	03-216-3211	富士通	59800	
COMSTAR 2424AT/5	5	—	—	○	○	○	○	○	○	—	03-798-7846	日本電気テクノマーケティング	66000	
AX2400c	5	—	○	○	○	—	—	○	—	○	03-220-0535	ヒューコム	150000	
AX2424c	5	—	○	○	○	○	○	○	—	○	03-220-0535	ヒューコム	180000	
AX2400cJ	5	—	○	○	○	○	○	○	—	—	045-433-1211	フジ・ディジタル・システムズ	180000	
CTS624	5	○	○	○	○	○	○	○	—	—	03-814-3081	共和理化	0	
AX2400cJ	5	—	○	○	○	○	○	○	—	—	03-341-2566	丸菱エレクトロニクス	215000	
PC2400c	5	—	○	○	○	—	—	○	—	○	03-220-0535	ヒューコム	215000	AX2400C のﾎﾞｰﾄﾞ型
HD2400c	5	—	○	○	○	—	—	○	—	○	03-220-0535	ヒューコム	460000	AX2400C 2台が一体
Multi-Tech MultiModem V32	5	○	○	○	○	○	○	○	—	—	045-441-8611	コア	360000	
Multi-Tech MultiModem 696EH	6	—	○	○	○	○	○	○	—	—	045-441-8611	コア	185000	
AX9624c	6	○	○	○	○	—	—	○	—	○	03-220-0535	ヒューコム	300000	19,200bps
AX9612c	6	—	○	—	○	—	—	○	—	○	03-220-0535		360000	19,200bps
AX9624cJ	6	—	○	○	○	○	○	○	—	—	045-433-1211	フジ・ディジタル・システムズ	360000	
AX9624cJ	6	○	○	○	—	—	—	—	—	—	03-341-2566	丸菱エレクトロニクス	430000	19,200bps
PC9624c	6	—	○	○	○	—	—	○	—	○	03-220-0535	ヒューコム	438000	AX9624C のﾎﾞｰﾄﾞ型
QX12K	7	—	○	○	○	○	○	○	—	○	03-220-0535	ヒューコム	198000	12,000bps
AX2472c	7	—	○	○	○	—	—	○	—	○		ヒューコム	240000	
QX30K	8	—	○	○	○	○	○	○	—	○	03-220-0535	ヒューコム	348000	30,000bps
QXV.32c	9	—	○	○	○	○	○	○	—	○	03-220-0535	ヒューコム	498000	30,000bps

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%niguts.nig.junet@relay.cs.net*
(*ddbj%niguts.nig.junet@relay.cs.net* 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%niguts.nig.junet@relay.cs.net.* 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.

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

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				

DDBJニュースレター及びマニュアル配布申し込み書

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 Sequence Data Submission Form (MS-DOS 2HD/2DD Floppy) [ ]

DDBJニュースレター [ ] 新規 [ ] 継続 [ ] 中止
[ ] 定期配布 _____部 [ ] 一時配布 _____部

DDBJ計算機利用の手引き [ ]

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) [ ]

DDBJ/EMBL/GenBank Feature Table: Definition [ ]

GenBank/EMBL/PIR CD-ROM Structuring Proposal [ ]

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データベース等利用申請書

年　　月　　日

国立遺伝学研究所長　殿

貴研究所のＤＮＡデータベース等利用について下記のとおり申請します。尚、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　　年　月　日～　　年　月　日

※ＩＤ_____ ※ユーザネーム_____

※パスワード _____　ローマ字_____

3 利　職　名 _____　氏　名 _____印

用　英訳 _____

申　所　属 _____

請　〒 _____ TEL ____ (_____) _____

者　英訳 _____

所在地 _____

4 利用

目的 _____

5 利用計算機名 ___ M380Q/UNIX　6 ディスク利用量　M380Q:_____MB

___ MicroVAXⅡ/VMS　VAX:_____MB

7 接続方法 ___電話 ___DDX-P (___第一種 ___第二種) 8 通信速度 ___1200 ___2400

9 支　職　名 _____　氏名 _____印

払

責　所　属 _____

任　〒 _____　TEL ____ (_____) _____

者　所在地 _____

10経　職　名 _____　氏名 _____印

理

責　所　属 _____

任　〒 _____ TEL ____ (_____) _____

者　所在地 _____

11利用　12支出 ___国立学校校費 ___附属病院校費 ___文部省科学研究費

見込額 _____円　科目 ___研究所校費 ___国立学校受託研究費 ___その他 (　)

※については記入しないで下さい。

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。尚、「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、大学院学生は［博士］［修士］の課程を記入して下さい。

　　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。尚、大学院学生は、研究科名、専攻まで記入して下さい。英訳をお付け下さい。

　　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。尚、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データベース等利用を必要とする研究テーマを記入して下さい。

5　利用計算機名　　利用する計算機名にチェックして下さい。不明な場合は記入しなくて結構です。

6　ディスク利用量　　ディスク利用量を記入して下さい。尚、長期保存のディスクは、最大10MBまでです。不明な場合は記入しなくて結構です。

7 接続方法及び 8 通信速度

　　該当する事項にそれぞれチェックして下さい。不明な場合は記入しなくて結構です。

9 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、その支払いに責任の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時は、［利用申請者に同じ］と記入して下さい。

10 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。たとえば、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［支払責任者に同じ］と記入して下さい。

11 利用見込額　　利用料金の見込額を記入して下さい。

12 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。

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／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	年　　月　　日	備考	

Mailing Addresses for Inquiries

If you have any inquiry, please send mails to the following addresses (...@niguts.nig.junet).

nig	about the nig system
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, gbsub, embl, and emblsub will be forwarded to them.

NOTE: On the nigvms, type "$ smtp ...@niguts.nig.junet".

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
E-mail: ddbj@niguts.nig.junet

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 (dbs)

(@niguts.nig.junet)