現代化学
CHEMISTRY TODAY

別　刷

東京化学同人

核酸データベースの現状と将来

宮　沢　三　造

生化学の分野で最も利用されているデータベースの一つが核酸の塩基配列のデータベースである．日，米，欧州共同でのデータの収集，提供が行われる一方，使いやすいデータベースを目指して，さまざまな改良も進んでいる．

はじめに

1970年代後半に始まるDNA塩基配列解析技術の進歩により，さまざまな生物種において多くの遺伝子がDNAレベルで解析されるようになり，DNA塩基配列の報告は指数関数的に増大した．このような状況の中で，1982年，欧州にEMBL Data Library（EMBL：ヨーロッパ分子生物学研究所），米国にGenBankがDNAデータバンクとして国の援助の下に設立された．また日本でも1986年，DNA Data Bank of Japan（DDBJ）が国立遺伝学研究所に設立された．現在これらのデータバンクは，重複を避け，共同でデータの収集，提供を行っている．

DNAデータベースは，塩基配列の相同性検索（homology search）および塩基配列の整列（sequence alignment）をおもな解析手段として，塩基配列の相同性の程度に基づく分子系統樹の作成や相同性を手がかりにしたDNAまたはタンパク質の機能予測，構造予測など，分子進化の研究ならびにDNAの遺伝情報解析に欠くことができない．近ごろでは新しいDNA塩基配列の解析ができた場合，配列の相同性検索は実験家のルーチンワークとなっている．このように生物学，医学，農学などの広範な分野にわたって研究上必要不可欠となったDNAデータベースの現状と将来について述べる．

1．DNAデータベースとは

現在DDBJ/EMBL/GenBankデータベースは，解析されたDNA断片ごとに図1aに示されるようなエントリー単位で管理され，フラットファイル（文字から成る行の単純な羅列）の形で配布されている．図1aはDDBJ/GenBankのエントリーの例である．EMBLのエントリーも書式は異なるが本質的に同様である．図1aで示されるように，1エントリーは，エントリー名，データを一意的に指定する受理番号，塩基配列の定義，配列が由来する生物種の生物分類，配列データの参照論文，注釈テーブル（feature table）と呼ばれるDNA塩基配列にコードされている遺伝情報についての記述，そしてDNA塩基配列などを記述した各種のレコードからなる．ここで示した注釈行は，図1bで示した配列情報に対応し，1990年の8月からDDBJ/EMBL/GenBankで新しく採用し始めた書式で表現してある．

従来の書式は，塩基配列情報に関して新知見が急速に明らかになりつつあるにもかかわらず，計算機処理が可能なように的確に表現できず，そのため注釈情報の計算機プログラムによる自動処理が困難であった．図1aで示した新書式は，このような従来の書式の欠点を克服する目的で，1986年以来EMBL/GenBank（1987年からはDDBJも参加）が共同研究し，1988年9月に完成させたものである．注釈テーブルは，feature keyと呼ばれるキーワードを用いて塩基配列の遺伝情報のタイプを記述する欄と，その情報に対応する配列断片を指定するlocation fieldから成る．遺伝子名，タンパク質名などの属性はqualifier（属性リスト）としてlocation fieldに記述される．表1，2で示されるように，複雑で多様な遺伝情報を記述すべく豊富なキーワード（feature key）と属性リスト（qualifier）をもっている．また新知見のための拡張性に優れていると同時に，計算機処理が可能な書式になっている．

2．DNAデータベースの現状

図2はGenBank，EMBLとDDBJデータベースに収集されている塩基数の変遷を示したものである．1985年ごろからGenBankとEMBLは重複がないようにデータ入力を分担し，互いに交換したデータを個々の書式でリリースしている．したがってGenBankとEMBLデータベースは実質的には同じ内

a)

```
LOCUS       HUMIGKVAC    1331 bp ds-DNA              PRI       15-MAR-1990
DEFINITION  Human Ig germline kappa light chain V-region (VkIII) gene, partial
            cds, clone Humkv328h5.
ACCESSION   M23090
KEYWORDS    immunoglobulin; immunoglobulin light chain; kappa-immunoglobulin;
            variable region; variable region subgroup VkIII.
SOURCE      Human peripheral blood granulocyte DNA, clone Humkv328h5, from
            patient Les.
  ORGANISM  Homo sapiens
            Eukaryota; Animalia; Metazoa; Chordata; Vertebrata; Mammalia;
            Theria; Eutheria; Primates; Haplorhini; Catarrhini; Hominidae;
            Homo; sapiens.
REFERENCE   1 (bases 1 to 1331)
  AUTHORS   Liu,M.-F., Robbins,D.L., Crowley,J.J., Sinha,S., Kozin,F.,
            Kipps,T.J., Carson,D.A. and Chen,P.P.
  TITLE     Characterization of four homologous L chain variable region genes
            that are related to 6B6.6 idiotype positive human rheumatiod factor
            L chains
  JOURNAL   J. Immunol. 142, 688-694 (1989)
  STANDARD  full staff_review
COMMENT     Draft entry and computer-readable sequence for [1] kindly submitted
            by P.P.Chen, 13-MAR-1989.
FEATURES             Location/Qualifiers
     CDS             join(676..724,894..>1189)
                     /note="Ig kappa-chain V-region precursor
                     /nomgen='IGKV' /map='2p12'
                     /hgml_locus_uid='LF0081F'"
                     /partial
     sig_peptide     join(676..724,894..904)
                     /note="Ig kappa-chain signal peptide"
     intron          725..893
                     /note="Ig, intron A"
     mat_peptide     905..>1189
                     /note="Ig kappa-chain"
                     /partial
     iDNA            1189..>1331
                     /note="Ig kappa intervening DNA"
     misc_signal     1192..1198
                     /note="7 mer recombination signal"
BASE COUNT      366 a    312 c    317 g    336 t
ORIGIN
        1 gtaccagtat tgtcacagtt acacagatat ggaaaccgag acacagggaa gttaagttac
       61 ttgatcaatt tcaagcaatc ggcaagccat ggagcatcta tgtcagggct gccaggacat
        ...
        ...
        ...
     1261 tcctttacag acagctagtg tggtggccac tcagttttag catctctgct ctatttggcc
     1321 attttggagc t
//
```

b) ヒト免疫グロブリン生殖系列κ軽鎖V領域

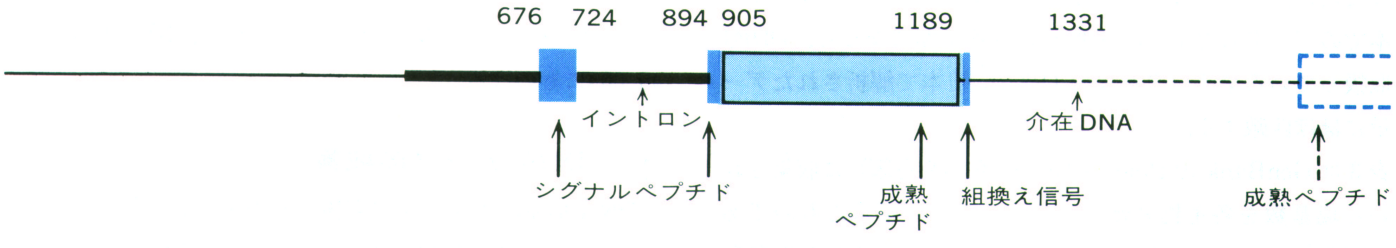

図1　GenBank/DDBJエントリーの例(a)，対応するDNA配列(b)
（EMBLも同様のフラットファイルを用いている）

表1　注釈キーワードとその階層性

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. repeat_unit
  a) repeat_unit
  b) LTR
  c) satellite

4. misc_binding
  1) primer_bind
  2) protein_bind

5. 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

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

表2　注釈テーブルにおける属性リストの例

属性リスト	例
/anticodon=(pos:,aa:)	/anticodon=(pos:34..36, aa:Phe)
/bound_moiety=	/bound_moiety="repressor"
/codon_start=	/codon_start=213
/direction=	/direction=LEFT
/function=	/function="essential for recognition of cofactor"
/gene=	/gene="ilvE"
/mod_base=	/mod_base=m5c
/note=	/note="a comment."
/phenotype=	/phenotype="erythromycin resistance"
/product=	/product="catalase"
/pseudo	/pseudo
/rpt_family=	/rpt_family="Alu"
/rpt_type=	/rpt_type=INVERTED
/rpt_unit=	/rpt_unit=Alu_rpt 1

表3　GenBank(Release65, 9/90)の内容

	レポート数	エントリー数	塩基数
霊長類	8735	6997	8434211
齧歯類	8427	7116	7251901
他の哺乳類	1653	1434	1748745
他の脊椎動物	2135	1769	2016044
無脊椎動物	3466	2915	3686353
植物	3240	2704	4122236
細胞小器官	1457	1193	1674157
細菌	5126	4015	6447923
構造RNA	1828	1533	428205
ウイルス	4546	3547	5963705
ファージ	825	522	633721
合成DNA	1108	1011	507546
無注釈データ	6441	4747	6264538
合計	48987	39533	49179285

容になっている．DDBJはデータをGenBank書式で入力し，GenBankとEMBLに提供している．DDBJが入力した量のみが図2に示されている．

データバンク間でのデータ入力の分担は，従来，学術雑誌単位に行われていたが，データバンクへのデータ提出を論文受理の条件とする学術雑誌が増加した結果，実験家から直接データが提供されるようになったので，1991年には，受理したデータは各データバンクで入力処理する体制に移行する計画である．DDBJはおもに日本で出版される雑誌を担当していたが，1989年末から直接提供されたデータの一部を入力処理するようになったため，世界で入力されたデータの約3％にすぎなかった日本でのデータ入力量が，1990年には約7％に達した．DDBJが取扱っている全データ量は約11％で，日本で解析されたデータ量にほぼ匹敵する．

表3はGenBankのRelease 65（1990年9月版）に収集されている塩基数を各生物分類カテゴリーごとに示したものである．図3はいくつかの生物種についてどのくらいDNA塩基配列が解析されているかを調べた結果である．比較すると図3にあげたような数少ない特定の生物種に集中して解析されてきたということがわかる．それらの生物種名をみると研究の方向がうかがえておもしろい．とはいっても，ヒトの場合で解析されたのは全ゲノムの約0.3％でしかない．大腸菌の場合でも約37％である．収集されている全塩基数は大腸菌ゲノムの約10倍に相当する約4900万塩基である．もちろん付加情報も含めれば，GenBank，EMBLともにそのデータベース量は塩基数の約2.5倍（110～130Mバイト）に達する．関連する論文も約49,000にものぼる．

3．DNAデータの収集

DNA塩基配列データが発表された論文から図1aで示されるようなデータを作成しようとすると，注釈データの作成に専門知識が要求されるため，時間と人手がかかり，近年のデータの増大にはとても追い付くことができない．そこでデータバン

クは現在，書式，feature key などの詳細を知らないでもデータ作成が可能なような，パーソナルコンピューターの上で稼働するデータ入力ソフトウェアを用意し，塩基配列を解析した実験家自身でデータ作成を行うシステムへ移行しつつある．GenBank により開発されたこのソフトウェアは，AuthorIn と呼ばれ，IBM/PC 版と Macintosh 版が GenBank から入手できる（PC 98 版は表 4 参照）．

データ収集におけるもう一つの問題は，いかに迅速かつ洩れなく収集できるかということである．データバンクの働きかけもあり，現在多くの学術雑誌が DNA データのデータバンクへの提出を論文受理の条件としている．このようなことが可能であるのも，DNA データベースが一部の研究者だけでなく，すべての研究者にとって必須だからである．

4．DNA データベースの利用

EMBL, GenBank の二つのデータベースは年 4 回，DDBJ は年 2 回最新版がリリースされている．日本では国立遺伝学研究所 DDBJ が，DDBJ，EMBL および GenBank のデータベースを磁気テープで配布している．GenBank と EMBL は CD-ROM でも配布している．計算機ネットワークを介してこれらのデータベースを入手することもできる．Internet と呼ばれる米国を中心に全世界に広がった計算機ネットワークに接続されている計算機では，“anonymous-ftp” というコマンドを使って，米国の “genbank. bio. net”，や日本の “flat. nig. ac. jp” という計算機にネットワーク経由でアクセスすれば，だれでもデータベースのファイルを自由に入手できる．また計算機間の電子メールを使用し．“netserve@embl. bitnet”（欧州），“retrieve@genbank. bio. net”（米国），“netserv@flat. nig. ac. jp”（日本）などの電子メールアドレスに，コマンドを記したメールを送付することにより，必要なエントリーを入手する

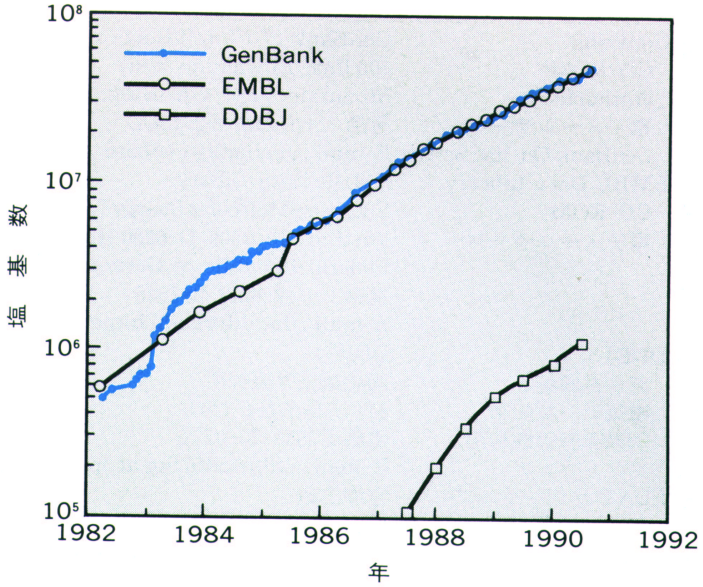

図 2　GenBank, EMBL, DDBJ 各データベースに収集されている塩基数の変遷
（DDBJデータベースはGenBank, EMBLに含まれている）

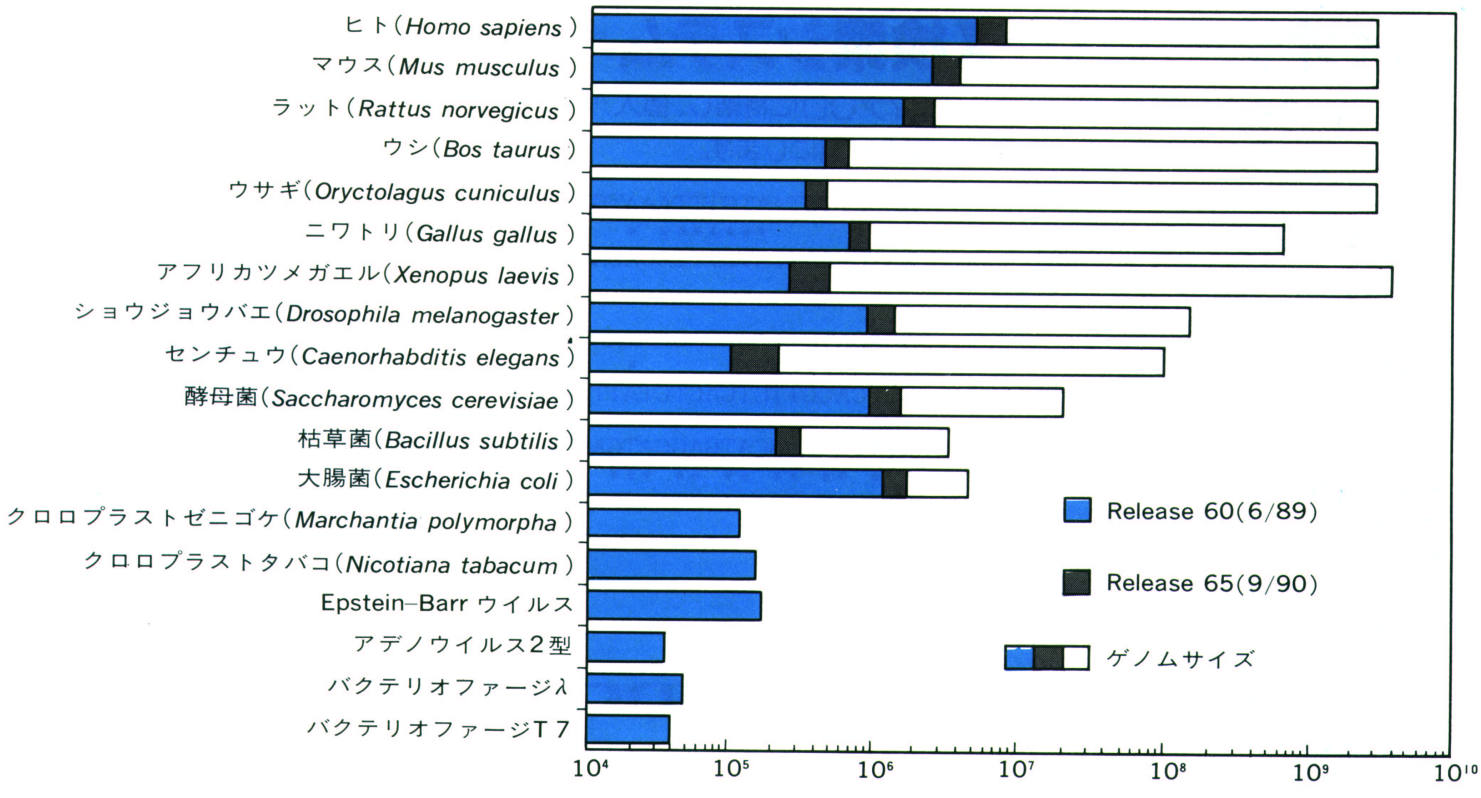

図 3　比較的解析が進んでいる生物種に関してそのゲノムサイズとGenBank（Release 60, 6/89；Release 65, 9/90）に収集されている塩基数

表 4　DNA 塩基配列データベースの問い合わせ先

● GenBank	GenBank c/o IntelliGenetics, Inc.
CD-ROM	700 East El Camino Real
AuthorIn	Mountain View, CA 94040, U.S.A.
電子メールサーバー	電話：+1-415-962-7364
GenBank On-line Service	E-mail : genbank@genbank.bio.net
● EMBL Data Library	EMBL Data Libary
CD-ROM	European Molecular Biology Laboratory
電子メールサーバー	Postfach 10.2209, D-6900 Heidelberg
	Federal Republic of Germany
	電話：+49-6221-387258
	E-mail : datalib@embl.bitnet
● DDBJ	DDBJ
データ収集	国立遺伝学研究所
磁気テープ配布	411 三島市谷田1111
DDBJ 計算機利用	電話：0559-75-0771
ソフトウエア	E-mail : ddbj@ddbj.nig.ac.jp
● FLAT ソフトウエア	宮澤三造
電子メールサーバー	国立遺伝学研究所
anonymous-ftp	E-mail : smiyazaw@flat.nig.ac.jp
NEC/PC98版 AuthorIn	

ことができる*．これらの計算機では各データバンクから電子メールで送られてくる最新のデータに基づきデータベースを毎日更新しているので，論文に発表されたばかりの最新のデータが入手できる．さらに，このうちの“netserv@flat. nig. ac. jp”（日本）では DDBJ/EMBL/GenBank のデータベースを検索できる．また GenBank は入力したデータを Usenet というネットワークの電子ニュースである Bionet ニュースグループに投稿し，全世界に流している．日本でも JUNET という電子メール・電子ニュースネットワークに接続された計算機であれば，これが利用できる．これらの利用法についてさらに詳しく知りたい方は，表 4 を参照してほしい．

DNA データベースを研究に利用するには，いうまでもなくデータベース検索と解析用のソフトウエアが必要である．それには，パーソナルコンピューター用のソフトウェアの利用や，国立大学共同利用大型計算機や DDBJ 計算機などの共同利用の計算機の利用，あるいは米国の GenBank On-line Service (GOS)の利用などが考えられる．もちろん各自が所有するワークステーションなどを用いて，研究者が作成し無料で入手できるソフトウェアを利用することもできる．代表的なミニコンピューターである VAX/VMS システムについては GCG, Ideas, PSQ/NAQ などのソフトウェアがある．UNIX システム用としては著者が FLAT というソフトウェアを配布している．

5．DNA データベースの将来

最近，人間の DNA の全塩基配列を解析しようとするヒトゲノム解析計画が，米国をはじめ日本でも発足した．大量の塩基配列の解析は DNA データベースに量的だけでなく質的変化を

＊　使用方法を知るには“help”と書いたメールを送る．

もたらすであろう．

現在データベースはエントリー単位で管理されている，いわゆるフラットファイルデータベースである．したがって複数のエントリーにまたがった修正は厄介である．たとえば，遺伝子名，タンパク質名などの変更である．データバンクとしてはこのような変更が容易にできるものがデータ管理の上から必要とされている．一方，データ検索に関しても現在は通常エントリー単位でなされているが，今後は遺伝子名，タンパク質名，map position（遺伝子地図上の位置）による塩基配列断片（断片的な塩基配列）の検索が必要とされるだろう．個々の配列断片からゲノム DNA 塩基配列を組立てるためには配列相互の重複部分の情報が必要である．また，ゲノムデータにおいては，配列に関する遺伝子地図データベースや物理的地図データベースとの相互参照は必須である．このような要求にこたえるべく，データバンクは，フラットファイルデータベースから関係データベース（relational date base）への移行を推進している．

関係データベースはテーブル（表）の集まりから構成されている．各テーブルは共通の値をとるコラム（欄）により互いに関係づけられる．もちろん遺伝情報は階層的構造もあるため関係データベースで表現するにはむずかしい点もあるが，現時点で利用できる管理システムとしては最適のものであろう．GenBank と EMBL は 1990 年，関係データベースに移行した．データベースを管理するソフトウェアは，GenBank は Sybase，EMBL は Oracle とそれぞれ異なるが，データベースへの追加変更に関するデータを一定の手順にしたがって交換することにより，実質的に同一のデータベースを構築する計画である．GenBank は，1990 年末，GenBank On-line Service で，関係データベースの公開を開始した．

今後も図 1 a のようなフラットファイルでのデータベースの配布は継続されるが，近いうちに関係データベースの配布も開始されるだろう．また関係データベースへの移行により種々の書式でデータを表現することがより容易になるので，いろいろな観点から編集されたデータベースが今後発表されるだろう．ゲノム解析計画が発足した生物種に関するゲノムデータベースは最も必要度の高いものの一つであろう．

参　考　書

1） 小谷正雄ほか 編，“蛋白質・DNA のデータバンクと情報解析”，蛋白質 核酸 酵素 別冊 No. 29，共立出版(1986)．

2） 宮澤三造，“ゲノム解析とデータベース”，実験医学，**8**，42，(1990)．

3） 宮澤三造，“DDBJ 計算機利用の手引”，遺伝研，DDBJ，(1988)．

4） 石田晴久，“コンピューターネットワーク”，現代化学 1990 年 11 月号“特集：コンピューターの化学・生化学”．