# 遺伝子－蛋白質解析システムGENASについて

久原 哲[1)]，高木利久[2)]，二村祥一[3)]，榊佳え[4)]，林 勝哉[1)]，松尾文碩[3)]
1)九州大学農学部農芸化学科，2)工学部情報工学科，3)大型計算機センター，4)遺伝情報実験施設

遺伝子－蛋白質解析システムGENASはAdbisをDBMSとしてEMBL Nucleotide Sequence Data Library，GenBank，NBRF Protein DataBankをデータベースとし応用プログラムの実行環境をもつデータベースシステムである。検索機能だけでなく，情報処理機能をもつことにより配列解析に有効であり，同時に個人の配列に対しても同様の解析ができることによりその応用範囲は広い。

GENASに登録されている応用プログラムは挿入欠出を含む文字列の高速検索や長大配列の２次構造推定などがあり有効な解析ができるようになっている。

霊長類の反復配列であるL1ファミリーの遺伝子，蛋白質の構造解析を行なった。その結果，相同性の高い領域が存在することが明らかとなり，同時にコードしている蛋白質がウィルスのRNA依存性DNAポリメラーゼと相同性があった。

"GENAS: A database system for nucleic acid and protein sequence analysis" (in Japanease)

Satoru KUHARA[1)], Toshihisa TAKAGI[2)], Syouichi FUTAMURA[3)], Yoshiyuki SAKAKI[4)], Katuya HAYASHI[1)], Fumihiro MATSUO[3)]

[1)]Department of Agricultural Chemistry, [2)]Department of Computer Science and Communication Engineering, [3)]Computer Center and [4)]Research Laboratory for Genetic Information, Kyushu University, 46-02, Fukuoka 812, JAPAN

A database system named GENAS,(GENe and protein Analyzing System) for computer analysis of sequence was constructed using a duductive DBMS called Adbis which manages integrated database GENEDB and application programs. GENAS enables us to retrieve any sequence data from EMBL nucleotide sequence data library, GenBank and NBRF protein sequence databank, and readily to analyze them (if necessary, together with private data).

Homology search program in application uses Wilber and Lipman's fast algorithm which allows insertion and deletion between a private sequence and database sequence. Second structure of large RNA sequence can be estimated by dynamic programing algorithm.

L1 family of repetitive DNA sequences in primate were analyzed homology among sequences using GENAS. Members showed a high homology among each other and long open reading frame in L1 family sequence has siginficant sequence homology to several RNA-dependent polymerase of virus.

## 1　はじめに

遺伝子，蛋白質は生体機能をつかさどる重要な生体高分子であり，古くから研究が行なわれ多くの研究成果が発表されている。その構成要素は，遺伝子ではアデニン（A），グアニン（G），チミン（T），シトシン（C）の4種であり，蛋白質では20種のアミノ酸である。1950年代に蛋白質ではポーリングによってα-ヘリックスという2次構造のモデルが，遺伝子ではワトソン-クリックによって遺伝子（DNA）の2重ラセン構造のモデルが出され，構造と機能についての研究がさかんとなるきっかけを作った。特にDNAの2重ラセン構造は遺伝子の複製つまり遺伝に重要な知見を与え遺伝から分子生物学への橋渡しを行なった。蛋白質については構造解析の手法が1950年代に確立され，蛋白質のアミノ酸配列が次々に決定されていた。しかしながら当時は遺伝子の構造解析の有効な手法がなく解析はあまり進まなかった。1970年代になり，サンガー法やマキサム-ギルバート法などの遺伝子構造解析手法の確立と遺伝子工学と呼ばれる遺伝子操作技術の普及とが相まって容易にかつ短時間で遺伝子の構造が決定できるようになった。これ以降は遺伝子の塩基配列の発表数は飛躍的に増加し，現在では蛋白質のアミノ酸配列として発表されるもののうち約90％が直接蛋白質から配列を決定したものでなく遺伝子の上で決定されたものという状況になっている。

この遺伝子，蛋白質の配列の急激な増加に際して，配列情報のデータベース化が提案された。蛋白質に関しては1950年代からジョージタウン大学のNational Biomedical Research Foundation（NBRF）の故Dayhoff女史らが中心となりアミノ酸配列を収集しAtlas of Protein Sequence and Structure[1]を出版し，1971年からはProtein Data Bank（NBRF蛋白質データバンク）となり計算機可読な形となっている。これに対し，遺伝子では1979年にデータベース化の動きが始まり，欧州では欧州分子生物学研究所（EMBL），米国ではNational Institute of Health（NIH）が中心となり遺伝子データベースを作り始めた。1982年には，EMBLがEMBL Nucleotide Sequence Data Library Release 1.0を米国ではLos Alamos国立研究所とBolt Beranek and Newman（BBN）社がGenBank Release 1.0を配布し始めた。

1980年代に入ってオンコジーンと呼ばれる発癌蛋白質をコードしている遺伝子[2]が発見されたが蛋白質の機能は不明であった。しかしデータベースとの相同性の検索により，それまでに知られていた蛋白質リン酸化酵素，増殖因子あるいはDNA結合蛋白質などに非常に類似していることが明らかとなり発癌機構の解明に一歩近ずくことになった。この結果，分子生物学の分野でもデータベースが盛んに利用されるようになってきている。

しかしながら，我国においては検索と処理機能を持つデータベースシステムが存在していなかった。そこで著者らは九州大学大型計算機センターのFACOM M-382 OSⅣ/F4上に遺伝子-蛋白質解析システムGENAS[3]を構築した。GENASは情報検索と情報処理の2つの機能を持ち，GenBank，EMBL Nucleotide Sequence Data LibraryとNBRFのProtein Data Bankをデータとし，データベース管理システムとして九州大学大型計算機センターが開発したデータベース統合支援システムAdbis[4]を使用した関係型データベースシステムである。

## 2　遺伝子の構造と蛋白質の発現機構

ここで遺伝子の構造と遺伝子にコードされている蛋白質の発現機構について簡単に述べる。この発現の機構は図1に示すように転写，プロセッシング，翻訳の3過程に分けられ，それぞれの過程は配列と密接な関係にある。

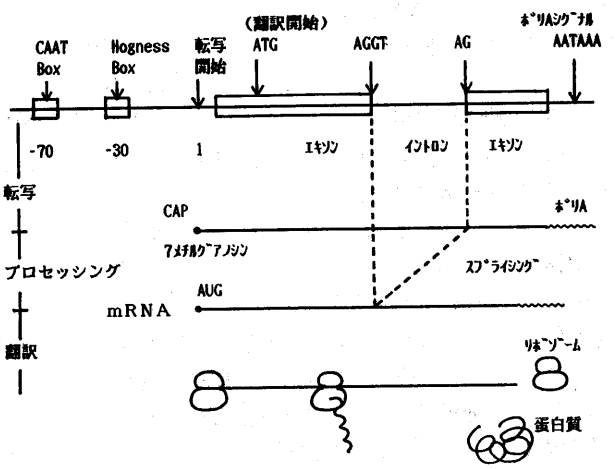



図1　蛋白質の発現機構

2－1　転写過程

発現の最初はRNAポリメラーゼと呼ばれる酵素により，DNAの情報がRNAへ転写される過程である。転写が起る部位の周辺には各遺伝子間で共通性の高い2つの配列がある。1つはHogness box あるいはTATA box と呼ばれるもので転写開始点の上流30塩基付近にあり次のような配列を持っている。

$$T_{90}A_{95}T_{90}A_{85}\begin{matrix}A_{56}\\T_{43}\end{matrix}T_{48}A_{90}A_{34}$$

ここでA,T,C,Gの下に付けた数字は出現頻度を示している[5]。このHogness boxのさらに約50塩基上流にCAAT boxと呼ばれる

$$GG\begin{matrix}C\\T\end{matrix}CAATCT$$

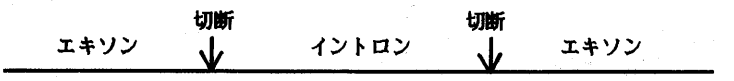
の共通な配列[6]があり，これらの共通配列は転写効率に影響することが知られている。これに対して転写の終了点の配列については不明であり，わずかに原核生物の転写終了の配列が明らかになっているに過ぎない。

2－2　プロセッシング過程

ポリメラーゼによる転写産物であるRNAは5′末端に7メチルグアノシン(CAP)，3′末端にはポリAが付加される。CAPが付加される位置はポリメラーゼによる転写開始に対応し，ポリAが付加される位置の10～20塩基上流にはAATAAAのポリAシグナルと呼ばれる共通配列が存在している。

真核生物では蛋白質をコードしている遺伝子はいくつかの部分に分かれて存在しているという特徴がある。蛋白質をコードしている部分をEXON（エキソン），していない部分をINTRON（イントロン）と呼ぶ。このような構造をとっているRNAを前駆体RNAと呼び，この前駆体RNAからINTRONが切りとられEXONのみから成る成熟メッセンジャーRNA（mRNA）ができる。この過程をスプライシングと呼ぶ。このスプライシング過程は蛋白質の配列に関係するものだけに正確さが要求され，その為にはINTRONとEXONの連結部位に何らかのシグナルが存在していると考えられ図2に示すような共通構造[7]が見つけられた。しかしながら

スプライシングの詳細な機構はいまだ不明な点が多い。




AAGGTAAGT . TTNCAGG

図2　スプライシング部位の共通配列

2－3　翻訳過程

プロセッシングを受けてできたmRNAは次にリボゾームと呼ばれる装置により蛋白質に翻訳される。その過程はまずメチオニンを示すAUGのmRNA配列にリボゾームが結合し，次に表1に示す3ヌクレオチド（コドン）に従ってアミノ酸に翻訳され，アミノ酸間でペプチド結合が形成され蛋白質の一次構造が完成する。

表1　遺伝コード表

| First position (5'-end) | Second position U | Second position C | Second position A | Second position G | Third position (3'-end) |
|---|---|---|---|---|---|
| U | Phe | Ser | Tyr | Cys | U |
| | Phe | Ser | Tyr | Cys | C |
| | Leu | Ser | Stop | Stop | A |
| | Leu | Ser | Stop | Trp | G |
| C | Leu | Pro | His | Arg | U |
| | Leu | Pro | His | Arg | C |
| | Leu | Pro | Gln | Arg | A |
| | Leu | Pro | Gln | Arg | G |
| A | Ile | Thr | Asn | Ser | U |
| | Ile | Thr | Asn | Ser | C |
| | Ile | Thr | Lys | Arg | A |
| | Met | Thr | Lys | Arg | G |
| G | Val | Ala | Asp | Gly | U |
| | Val | Ala | Asp | Gly | C |
| | Val | Ala | Glu | Gly | A |
| | Val | Ala | Glu | Gly | G |

3　遺伝子－蛋白質解析システムGENAS

3－1　データベース

遺伝子の塩基配列データベースとしては欧州ではEMBLがデータ収集配布を行なっているEMBL Nucleotide Sequence Data Libraryと米国ではNIHによりLos Alamos国立研究所がデータ収集を行ないBBN社がデータ配布を行なっているGenBank，およびNBRFがデータ収集配布を行なっているNBRF Nucleotide Data Bankの3つがある。これらのデータベースは1982年から公共データベースとして無料で磁気テープあるいはフロッピーディスクによる配布が開始されており版権などの制限はない。GenBank，NBRFは年4回またEMBLは年2回更新されている。

蛋白質のアミノ酸配列のデータベースとしてはNBRFのDayhoffらがAtlas of Protein Sequence and Structureを出版しており，1983年からは公共データベースNBRF Protein Data Bankとして無料で配布されている。これらのデータベースの増加は塩基配列に関しては年間150万塩基以上になっている。

これら3つのデータベースはそれぞれ異なった書式を用いて書かれており，その例を図3に示す。EMBLを例にとって構造を説明すると1つの遺伝子の情報は1つのエントリーとして格納されている。各エントリーはまずID行で遺伝子を表わすコード名，AC行でアセッション番号，DT行で入力された日付，DE行で遺伝子の定義，KW行でキーワード，OS行で生物種，OC行で分類，RN行で文献番号，RT行で文献題目，RA行で著者名，RL行で文献名，CC行で注釈，FH行でfeature tableの見出し，FT行でfeature table，SQ行で配列の見出し，その次に配列，最後に//行で終了する順になっている。特に分子生物学の研究者が必要とする情報はFT行に示されているfeature tableであり，各情報の位置とコードがKey，From，To，Descriptionの順で示されている。

著者らのシステムでは，これらのデータベースの中で塩基配列としてEMBL，GenBankおよび蛋白質のアミノ酸配列としてNBRF Protein Data Bankを基にして新たに関係型データベースGENEDBを構築している。GENASを構築する上

```
ID   A1MVRNA2  STANDARD; RNA; 2593 BP.
XX
AC   X01572;
XX
DT   30-JAN-1986  (AUTHOR REVIEW)
DT   17-JUL-1985  (FIRST ENTRY)
XX
DE   ALFALFA MOSAIC VIRUS (A1M4) RNA 2
XX
KW   UNIDENTIFIED READING FRAME.
XX
OS   ALFALFA MOSAIC VIRUS
OC   VIRIDAE; SS-RNA NONENVELOPED VIRUSES; ALFAMOVIRUS.
XX
RN   1    (BASES 1-2593; ENUM. 1 TO 2593)
RA   CORNELISSEN B.J.C., BREDERODE F., VEENEMAN G.H., VAN BOOM J.H.,
RA   BOL J.F.;
RT   "COMPLETE NUCLEOTIDE SEQUENCE OF ALFALFA MOSAIC VIRUS RNA 2";
RL   NUCL. ACIDS RES. 11:3019-3025(1983).
XX
CC   DATA KINDLY REVIEWED (30-JAN-1986) BY J.F. BOL
XX
FH   KEY         FROM    TO       DESCRIPTION
FH
FT   TRANSCR       1    2593       A1MV RNA 2
FT   CDS          55    2424       UNIDENTIFIED READING FRAME (AA 1-790)
XX
SQ   SEQUENCE  2593 BP; 736 A; 533 C; 547 G; 777 U.
     GUUUUUAUCU UUUCGCGAUU GAAAAGAUAA GUUUUUCAGU UUAAUCUUUU CAAUAUGUUC
     ACUCUUUUGA GAUGUCUCGG AUUCGGUGUU AAUGAACCUA CUAACACUUC CUCAUCAGAG
     UAUGUUCCCG AGUAUUCCGU UGAAGAGAUU UCCAACGAAG UCGCUGAACU CGAUUCAGUG
     GAUCCAUUAU UCCAAUGUUA CAAACAUGUU UUUGUAUCAU UGAUGCUCGU GAAAAGAUG
     ACUCAAGCUG CCGAAGACUU CCUCGAGAGU UUUGGGGGAG AAUUCGAUAG CCCUUGUUGU
     AGGGUUUACC GUCUUUAUAG ACAUUUUGUU AAUGAAGACG AUGCACCCGC UUGGGCCAUA
     CCGAAUGUCG UGAAUGAAGA UUCUUACGAC GAUUAUGCCU ACCUCCGAGA GGAGUUAGAU
     GCCAUAGACA GCUCUUUUGA GUUGCUAAAC GAAGAGCGUG AGUUAUCGGA AAUUACGGAC
     AGACUCAACG CUUUAAGAUU UUUCCCUGUU UCUAAAACAG AAGCGCUACC AGUGGCGAAU
     GUCCAAGAGG UCAAACUCAU UUCUGAGACA UACCAGUUAU UGAUGACCUU UAUUAACUAC
     UCUGACGAGA AUAUUCCGUC UGAAAUGCCC GCACCAUUAC UGGAUGAGUU GGGGAUGUUA
     CCGGAGGAAC UUGGACCUCU GAAUGAAAUU GAAGACAUUA AGCCGGUGGC GGCUCCAAUC
     ACAUUACUAU CUGAGUUUAG AGCCUCAGAU AAUGCUAAGC CACUCGACAU AGUCGAAAUC
     AUUCCAGACG UAAGUCCGAC GAAACCUUAU GAAGCCGUCA UAUCAGGUAA UGAUUGGAUG
     ACGUUGGGGA GGAUCAUACC UACCACUCCC GUUCCUACCA UAAGGGAUGU CUUCUUCUCU
     GGUCUUUCUC GGCACGGAUC GCCGGAAGUG AUCCAGAAUG CUCUUGAUGA AUUUCUUCGG
     CUCCAUCAUU CAAUUGAUGA UAAGUAUUUU CAAGAAUGGG UUGAAACCUC AGAUAAAUCU
     CUCGAUGUCG AUCCAUGUCG AAUCGAUCUG AGUGUUUUCA ACAACUGGCA GUCUUCGGAA
     AACUGCUAUG AACCUCGGUU UAAAACCGGU GCAUUAUCCA CACGUAAGGG CACUCAAACU
     GAAGCCCUAU UAGCGAUAAA GAAACGUAAU AUGAAUGUGC CUAACCUGGG GCAGAUUUAU
     GACGUGAAUU CUGUUGCUAA UUCCGUGGUU AAUAAGCUCU UAACAACUGU UAUAGAUCCU
     GAUAAGCUGU GCAUGUUUCC AGAUUUCAUA UCUGAGGGUG AAGUUUCGUA UUUCCAGGAC
     UAUAUAGUUG GGAAGAAUCC CGACCCUGAA UUAUAUUCAG AUCCUCUAGG UGUUCGUUCC
     AUCGAUAGCU AUAAACACAU GAUUAAAUCC GUGUUAAAGC CCGUUGAAGA UAAUUCUCUG
     CACCUAGAAC GGCCGAUGCC AGCAACCAUA ACAUACCAUG AUAAAGAUAU CGUGAUGUCA
     UCUUCACCAA UUUUUUUGGC UGCUGCUGCC CGCUUGAUGU UAAUCUUAAG AGAUAAGAUA
     ACCAUACCAA GCGGAAAAUU CCAUCAUUUG UUUUCCAUCG AUGCUGAAGC CUUUGAUGCA
     AGUUUCCAUU UUAAAGAGAU AGACUUUUCG AAGUUUGACA AAAGUCAAAA UGAGUUGCAU
     CACUUGAUCC AGGAAAGGUU UCUGAAAUAC UUAGGUAUAC CCAACGAAUU UCUAACCUUA
     UGGUUUAAUG CGCAUAGAAA AUCCCGAAUC UCAGAUUCGA AGAAUGGCGU UUUUUUAAAC
     GUCGAUUUCC AACGUCGUAC UGGAGAUGCG CUCACGUACU UGGGAAACAC AAUAGUGACA
     UUAGCUUGUC UGUGUCACGU GUAUGACUUG AUGGACCCAA AUGUGAAAUU CGUUGUUGCU
     UCCGGUGAUG AUUCAUUGAU AGGCACUGUG GAGGAAUUAC CAAGAGAUCA AGAGUUUCUU
     UUCACGACUC UUUUUAAUCU UGAAGCAAAG UUUCCUCAUA ACCAGCCUUU CAUAUGCAGU
     AAGUUUUUGA UUACUAUGCC CACUACAAGU GGAGGCAAAG UUGUCCUGCC GAUACCGAAU
     CCAUUGAAAC UCCUCAUACG CUUGGGUUCG AAGAAAGUCA AUGCCGAUAU AUUCGAUGAA
     UGGUAUCAAU CUUGGAUUGA UAUAAUUGGU GGUUUUAACG ACCACCAUGU CAUCCGAUGC
     GUUGCCGCGA UGACAGCACA UAGGUAUCUC AGAAGACCGA GUUUAUACCU AGAAGCUGCU
     UUGGAAUCCC UAGGUAAGAU CUUCGCUGGU AAGACCUUGU GUAAGGAAUG CCUCUUUAAU
     GAGAAGCACG AGUCUAAUGU AAAAAUUAAG CCUCGUAGAG UGAAAAAAUC CCACUCGGAU
     GCCAGGUCAA GGGCACGCCG AGCUUGAUGU UUUCUUGACA UAAGUCAAAU UGCCAACCUC
     CACUGGGUGG GUCAAGGUUG AGGUAUAGAA UCCUAUUCGC UCCUGAUAGG AGAAAUUCUA
     UAUUGCUUAU AUAUGUGCUU ACGCACAUAU AUAAAUGCUC AUGCAAAACU GCAUGAAUGC
     CCCUAAGGGA UGC
//
```

```
LOCUS       APEMBA2M      549 BP    MRNA            ENTERED   01/07/85
DEFINITION  APE (CHIMPANZEE) ALPHA-2-GLOBIN MRNA.
ACCESSION   X00227
KEYWORDS    ALPHA-2-HEMOGLOBIN; ALPHA-GLOBIN; GLOBIN; HEMOGLOBIN.
SOURCE      CHIMPANZEE RETICULOCYTE (RBC), CDNA TO MRNA, CLONE PCH-ALPHA-1B.
  ORGANISM  PAN TROGLODYTES
            EUKARYOTA; METAZOA; CHORDATA; VERTEBRATA; TETRAPODA; MAMMALIA;
            EUTHERIA; PRIMATES.
REFERENCE   1  (BASES 1 TO 549)
  AUTHORS   LIEBHABER,S.A. AND BEGLEY,K.A.
  TITLE     STRUCTURAL AND EVOLUTIONARY ANALYSIS OF THE TWO CHIMPANZEE
            ALPHA-GLOBIN MRNAS
  JOURNAL   NUCL ACID RES 11, 8915-8929 (1983)
COMMENT     1  COMPARES THE SEQUENCES OF APE AND HUMAN ALPHA-1 AND ALPHA-2
            GLOBIN MRNA.  EVIDENCE OF A RECENT GENE CONVERSION IN THE HUMAN
            ALPHA-GLOBIN COMPLEX AND A MARKED HETEROGENEITY IN THE RATE OF
            STRUCTURAL DIVERGENCE WITHIN THE ALPHA-GLOBIN GENE WERE FOUND.
FEATURES        FROM  TO/SPAN     DESCRIPTION
   PEPT           21       449    ALPHA-2-GLOBIN
  SITES
   MRNA/MRNA       1         0    A2-GLOBIN MRNA UNSEQUENCED/SEQUENCED
   ->PEPT         21         1    A2-GLOBIN CDS START
   REFNUMBR       24         3    NUMBERED CODON 1 IN  1
   PEPT<-        449         1    A2-GLOBIN CDS END
   MRNA/MRNA     550         0    A2-GLOBIN MRNA SEQUENCED/UNSEQUENCED
BASE COUNT       96 A    203 C    151 G     99 T
ORIGIN      160 BP UPSTREAM OF TAQI SITE.
          1 AGACTCAGAA AGAACCCACC ATGGTGCTGT CTCCTGCCGA CAAGACCAAC GTCAAGGCCG
         61 CCTGGGGTAA GGTCGGCGCG CACGCTGGCG AGTATGGTGC GGAGGCCCTG GAGAGGATGT
        121 TCCTGTCCTT CCCCACCACC AAGACCTACT TCCCGCACTT CGACCTGAGC CACGGCTCTG
        181 CCCAGGTTAA GGGCCACGGC AAGAAGGTGG CCGACGCGCT GACCAACGCC GTGGCGCACG
        241 TGGACGACAT GCCCAACGCG CTGTCCGCCC TGAGCGACCT GCACGCGCAC AAGCTTCGGG
        301 TGGACCCGGT CAACTTCAAG CTCCTAAGCC ACTGCCTGCT GGTGACCCTG GCCGCCCACC
        361 TCCCCGCCGA GTTCACCCCT GCGGTGCACG CCTCCCTGGA CAAGTTCCTG GCTTCTGTGA
        421 GCACCGTGCT GACCTCCAAA TACCGTTAAG CTGGGGACTC GGTGGCCGTT CCTCCTGCCC
        481 GCTGGGCCTC CCAACGGGCC CTCCTCCCCT CCTTGCACCG GCCCTTCCTG GTCTTTGAAT
        541 AAAGTCTGA
//
```

```
>CCHU       104 AMINO ACIDS                          0     1     22      3 H
CYTOCHROME C - HUMAN, CHIMPANZEE, AND RHESUS MACAQUE                      T+
SPECIES: HOMO SAPIENS (HUMAN)) (CHIMPANZEE)) MACACA MULATTA (RHESUS       C+
    MACAQUE)                                                              C
MATSUBARA, H., AND SMITH, E.L.; J. BIOL. CHEM. 238, 2732-2753, 1963       R+
    (HUMAN, SEQUENCE WITH EXPERIMENTAL DETAILS)                           R
MATSUBARA, H., AND SMITH, E.L.; J. BIOL. CHEM. 237, PC3575-PC3576, 1962   R+
    (HUMAN)                                                               R
    65-LEU IS FOUND IN 10% OF THE MOLECULES IN POOLED PROTEIN.            C+
NEEDLEMAN, S.B., AND MARGOLIASH, E.; UNPUBLISHED RESULTS, 1966, CITED BY  R+
    MARGOLIASH, E., AND FITCH, W.M., ANN. N.Y. ACAD. SCI. 131, 359-381,   R
    1968; AND NEEDLEMAN, S.B., SUBMITTED TO THE ATLAS, OCTOBER 1968       R
    (CHIMPANZEE, COMPOSITIONS OF CHYMOTRYPTIC PEPTIDES)                   R
    CHIMPANZEE CYTOCHROME C APPEARS TO BE IDENTICAL WITH HUMAN.           C+
ROTHFUS, J.A., AND SMITH, E.L.; J. BIOL. CHEM. 240, 4277-4283, 1965       R+
    (RHESUS MACAQUE, COMPOSITIONS OF CHYMOTRYPTIC PEPTIDES AND SEQUENCES  R
    OF RESIDUES 55-61 AND 68-70)                                          R
    THIS SEQUENCE DIFFERS FROM THAT SHOWN IN HAVING 58-THR.               C+
THE AMINO END IS ACETYLATED.                                              C+
CYS-14 AND CYS-17 COVALENTLY BIND HEME; HIS-18 AND MET-80 ARE THE AXIAL    C+
    LIGANDS OF THE HEME IRON.                                             C
SUPERFAMILY: CYTOCHROME C                                                 C+
KEYWORDS: MITOCHONDRION; ELECTRON TRANSPORT; RESPIRATORY CHAIN;           C+
    OXIDATIVE PHOSPHORYLATION; HEME; ACETYLATION                          C
  G D V E K G K K I F I M K C S Q C H T V E K G G K H K.T G P N L H G L F S
  G R K T G Q A P G Y S Y T A A N K N K.G I I W G E D T L M E Y L E N P K S
  K Y I P G T K M I F V G I K K K E E R A D L I A Y L K.K A T N E         S
```

図3　データの書式
左：EMBL，右上：GenBank，右下：NBRF

2)関係のキーとして，各エントリーを一意に識別するためRF（エントリー参照番号）を定義した。このRFの範囲により EMBL，GenBank，NBRFの3つのデータを区別することにした。RFをキーとして四4に示すGENED3の関係を作った。特にSEQ(配列)については長大な配列（最大17万塩基対）が発表されており，その格納法に問題があるがGENASでは長さ500の固定長形式をとり，その個数を表わすSNを新しく作っている。その為SQについては，RF×SNの直積により SEQ を表わすことになる。またOC（分類）については，LVL（分類のレベル）とTAX（分類のエントリー）あるびRFの直積により表わすことにし

```
DATABASE(GENEDB)
        L('EMBL + GENBANK + PROTEIN DATABASE')
*
DEF(REF < RF*ID)
        L('RFERENCE CODE')
*
DEF(RNA < RF)
        L('RNA OR CDNA')
*
DEF(FR < RF)
        L('FRAGMENT')
*
DEF(DE : RF -> DEF)
        L('DEFINITION')
*
DEF(OS < RF*GEN*SPEC)
        L('ORGANISM')
*
DEF(OC < RF*LVL*TAX)
        L('ORGANISM CLASSIFICATION')
*
DEF(HS < RF*GEN*SPEC)
        L('THE MOST COMMON HOST SPECIES')
*
DEF(HC < RF*LVL*TAX)
        L('HOST CLASSIFICATION')
*
DEF(BIB : RF*SN -> BI)
        L('BIBLIOGRAPHIC INFORMATION')
*
DEF(KY < RF*KH)
        L('KEYWORD')
*
DEF(AUTH < RF*AU)
        L('AUTHOR')
*
DEF(DC < RF*DS)
        L('DATA COMPILE')
*
DEF(YEAR < RF*YR)
        L('YEAR')
*
DEF(FN : RF -> NFTE)
        L('NUMBER OF FEATURE TABLE ENTRIES')
*
DEF(FT : RF*FTEN -> KEY*FROM*TO*DESCR)
        L('FEATURE TABLE')
        I(KEY,FROM,TO)
*
```

```
DEF(SH : RF -> A*C*T*G*U*Y*R*AB)
        L('SEQUENCE HEADER')
*
DEF(LN : RF -> LEN)
        L('LENGTH OF SEQUENCE')
*
DEF(SQ : RF*SN -> SEQ)
        L('SEQUENCE')
*
DEF(RF)      KEY
             T(I4)    L('RFERENCE CODE')
*
*** PROTEIN IF RF > 1,000,000,000; OTHERWISE NUCLEOTIDE.
*
DEF(ID)      T(C10)   L('SHORT IDENTIFICATION')
*
DEF(GEN)     T(C100)  L('GENUS')
DEF(SPEC)    T(C50)   L('SPECIES')
*
DEF(LVL)     T(I2)    L('LEVEL IN TAXONOMIC TREE')
DEF(TAX)     T(C50)   L('ENTRY IN TAXONOMIC CLASSIFICATION')
*
DEF(DEF)     T(C600)  L('DEFINITION WRITTEN IN ENGLISH')
DEF(BI)      T(C2000) L('BIBLIOGRAPHIC INFORMATION')
DEF(KH)      T(C50)   L('KEY WORD')
DEF(AU)      T(C50)   L('AUTHOR')
DEF(DS)      T(C8)    L('DATA SOURCE')
DEF(YR)      T(I2)    L('YEAR')
*
DEF(NFTE)    T(I2)    L('NUMBER OF FEATURE TABLE ENTRIES')
DEF(FTEN)    T(I2)    L('FEATURE TABLE ENTRY NUMBER')
DEF(KEY)     T(C10)   L('FEATURE TABLE KEY NAME')
DEF(FROM)    T(I4)    L('ENDPOINT SPECIFICATION')
DEF(TO)      T(I4)    L('ENDPOINT SPECIFICATION')
DEF(DESCR)   T(C250)  L('MORE INFORMATION ABOUT FEATURE')
*
DEF(A)       T(I4)    L('NUMBER OF ADENINE BASES')
DEF(C)       T(I4)    L('NUMBER OF CYTOSINE BASES')
DEF(T)       T(I4)    L('NUMBER OF THYMINE BASES')
DEF(G)       T(I4)    L('NUMBER OF GUANINES')
DEF(U)       T(I4)    L('NUMBER OF URACIL BASES')
DEF(Y)       T(I4)    L('NUMBER OF PYRIMIDINE BASES')
DEF(R)       T(I4)    L('NUMBER OF PURINE BASES')
DEF(AB)      T(I4)    L('NUMBER OF ANY BASES')
*
DEF(LEN)     T(I4)    L('LENGTH OF SEQUENCE')
*
DEF(SN)      T(I4)    L('SEQ NUMBER')
DEF(SEQ)     T(C500)  L('NUCLEOTIDE/PROTEIN SEQUENCE')
```

図4　データ定義

た。最も重要であるFT(feature table)については NFTE(feature tableのエントリー数)を作りRF×FTENの直積をドメインとしてKEY×FROM×TO×DESCRの直積をコドメインとする関係とした。最後に3つのデータを区別するためDC(データコンパイル)の関係を作った。これらの関係の中で転置ファイルを作るものはFTのみとした。この定義に従って3つのデータより関係型データベースGENEDBを作った。

3-2　データベースシステム

前述のGENEDBをデータベースとしてGENASを作る上で次の2点に重点をおいた。

①情報の検索機能を持つ。但し検索手法は簡単であることが望ましい。

②情報の処理機能を持つ。特にデータが一次構造(配列)であり，共通配列とはいうものの変異があり配列の情報を得るには単純な文字列の検索だけでは不十分であり種々の処理を行なう必要がある。この2点を検討しGENASのモードを図5に示すように作った。

3-2-1　情報検索

まず情報の検索機能については，分子生物学の分野の研究者は計算機を使った経験があまりないので，Adbisの推論機能を直接用いる検索より，Adbisの情報検索型コマンドを使用することを考えた。次に検索を考えると，その検索のほと

んどがキーワードによる検索でありその他の検索，例えば詳細な分類による検索などはほとんど行なわないと考えられるのでGENASでは情報検索型コマンドで検索可能な領域はKW(キーワード)，AU(著者名)，ID(コード名)とし，キーワードに重点をおくことにした。基となるデータのキーワードでは不足なので前述のDE行，KW行，RT行，OS行から単語を切り出しすべてキーワードとした。

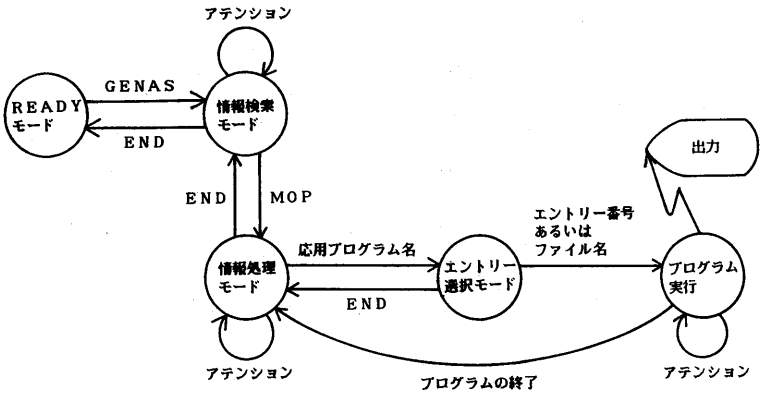



図5　GENASのモード遷移図

3-2-2　情報処理機能

遺伝子あるいは蛋白質解析のデータベースシステムでは一般の学術情報データベースとは異なり，前述したように構造と機能の関係が一意的に決定されておらず不確定な要素が多いため，定形の処理だけではなく多種の処理ができる環境を持たなければならない。

応用プログラムとのインタフェース：　Adbisのコマンド呼び出し機能を使用しMOPと呼ばれるコマンドプロシジャーを経由して応用プログラムを起動している。ただし応用プログラムの多様性および各応用プログラムのデータ要求の多様性からMOPの中で応用プログラムの選択を行ない，必要なデータあるいは環境の設定を行なっている。

表2　GENASの応用プログラム

| |
|---|
| 統計処理 |
| 塩基の出現頻度<br>コドンの出現頻度<br>類似性の統計処理 |
| 一次構造の文字列解析 |
| 相同性の検索<br>遺伝子からアミノ酸への翻訳<br>アミノ酸から遺伝子への逆翻訳<br>任意の文字列の検索 |
| 二次構造の予測 |
| 蛋白質の二次構造の予測<br>遺伝子の二次構造の予測 |

応用プログラム：　現在GENASに登録されている応用プログラムの一覧を表2に示した。応用プログラムの実行環境では特に個人が所有している配列に対してもデータベース中の配列と同様な処理ができるようにした。処理のダイアグラムを図6に示した。

分子生物学の分野では，先に紹介したようにオンコジーンの発見以来，実験で得た配列についてはデータベースの配列と相同性を検索することが一般的になってきている。この相同性の検索では文字列の完全一致ではなく文字列の不一致あるいは挿入欠失を含んだ一致を探さなければならない。不完全一致の相同性のプログラムについてはWilber，Lipmanらが開発したアルゴリズム[8]がよく用いられている。遺伝子の2次構造の推定にはダイナミックプログラミングの手法が用

いられ計算時間の短縮を行なっているものがある[9]。

3－3　実際例（霊長類に存在する反復配列L1ファミリーの一次構造の比較）[10]

著者らのグループは高等動物の遺伝子の特徴の1つである反復配列の1つであるL1ファミリーの6個（L1Nc1～L1Nc6）の塩基配列を決定しGENASによる一次構造の解析を行なった。その結果図7に示すようにL1ファミリーの間では90％以上の相同性があり，その中のL1Nc4に

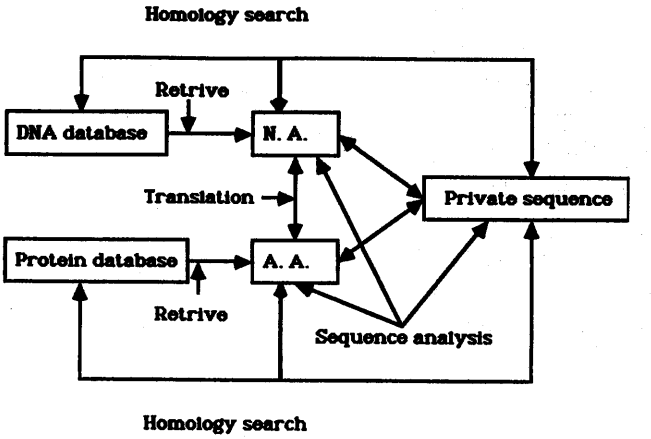



図6　処理のダイアグラム

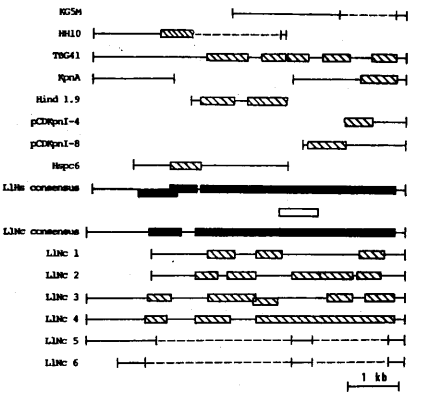



図7　L1ファミリーの構造と共通配列

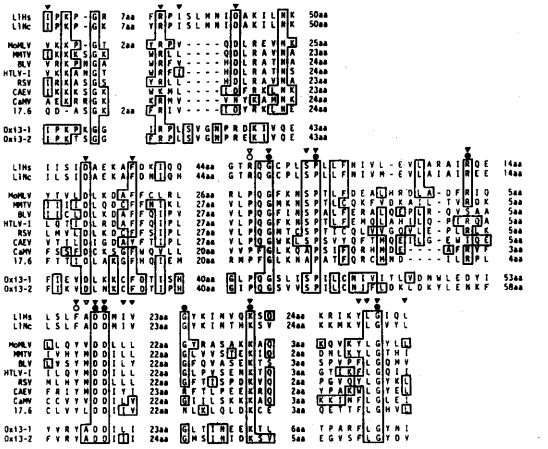


図8　L1のアミノ酸配列とウィルスのRNA依存性DNAポリメラーゼの相同性

は927アミノ酸残基から成る蛋白質がコードされている領域があることが明らかとなった。特に蛋白質をコードしている領域については他の種のL1ファミリーと65％以上の相同性があった。蛋白質をコードしている領域の相同性からみてL1ファミリーは進化的に保存されている蛋白質から誘導されてきたものであると考えられる。

次に，この保存されている蛋白質について検討した結果，図8に示すように，ウィルスのRNA依存性DNAポリメラーゼと部分的に強い相同性があることが明らかとなった。もし，コードしている蛋白質に逆転写活性があればL1ファミリーの分散の原因を考える上で重要な知見となるものである。

## 参考文献


1) Dayhoff,M.O., et al.(ed.) :Atlas of protein sequence and structure, Vol.1, National Biomedical Research Foundation, Silver Spring, Md.(1965)

2) Czernilofsky,A., et al. : Corrections to the nucleotide sequence of src gene of Rous sarcoma virus, Nature, Vol.301, pp. 736-738 (1981)

3) Kuhara,S., et al. :GENAS: A database System for Nucleic Acid Seuence Analysis, Nucleic Acids Res., Vol.12, pp.89-99 (1984)

4) 松尾文碩,二村祥一, 高木利久 :推論関係型データベース管理システムAdbis, 情報処理学論文誌, Vol.124, pp.249-255 (1983)

5) Lewin,B. :"GENES", Jhon Wiley & Sons, New York (1983)

6) Proudfoot,N.J., et al. :Structure of the Genes that do not Rearrange, Science, Vol.209, pp.1329-1336 (1980)

7) Lewin,B. :Alternatives for splicing: Recognizing the End of Introns, Cell, Vol.22, pp.324-326 (1980)

8) Wilber,W.J. and Lipman,O,J. :Rapid Simularity Searchs of Nucleic acid and Protein Data Bank, Proc. Natl. Acad. Sci. U.S.A., Vol.80, pp.726-730 (1983)

9) Zuker,M. and Stiegler,P. :Optimal Computer Foldinf of Large RNA Sequence using Thermodynamics and Auxiliary Information, Nucleic Acids Res., Vol.9, pp.133-148 (1981)

10) Hattori,M. et al. :L1 family of repetitive DNA sequences in primates may be derived from a sequence encoding a reverse transcriptase related protein, Nature, Vol.321, pp.625-628 (1986)