情　報　学　基　礎 33－3
（1994．5．18）

ヒトゲノム解析とGDB（ゲノムデータベース）

諏訪　秀策*、鈴木　政彦*、平川　美夏*、神田　利彦*、清水　信義**

*日本科学技術情報センター　技術開発部
**慶応義塾大学　医学部分子生物学教室

国際的なヒトゲノム解析計画が進行する中で、研究者にとってヒト遺伝子地図情報を集積したデータベースGDBの重要性が増している。GDBは国際的協調のもとに作成・維持されている。平成６年２月にGDB日本ノードが開設され、GDBのオンライン提供が始まった。本文では、最近の概況を述べ、若干の考察を加える。

Human Genome Project and GDB (Genome Data Base)

Shusaku Suwa*, Masahiko Suzuki*, Mika Hirakawa*, Toshihiko Kanda*, Nobuyoshi Shimizu**

*The Japan Information Center of Science and Technology
5-2, Nagata-cho 2 chome,Chiyoda-ku,Tokyo 100, Japan
**Keio University School of Medicine
35 Shinanomachi,Shinjyuku-ku, Tokyo 160, Japan

International human genome project has been progessing rapidly. The GDB, human genome data base, is becoming important more and more for researchers. The GDB is produced and maintained on international collaboration basis.

The GDB Japan node opened in Feb. 1994, and the GDB online services started. This paper describes recent status of the GDB and discusses some points.

1. はじめに

国際的に展開されているヒトゲノム解析計画（Human Genome Project）で重要な位置を占めている遺伝子地図のデータベースGDB（Genome Data Base）の最近の状況を報告し、若干の考察を行なう。

2. ヒトゲノム解析

生命体の設計図は細胞の中の染色体に折り込まれている。ヒト（人間）の染色体は1番から22番までの常染色体と、X、Yの性染色体がある。これらに遺伝情報が書き込まれており、全体のことをゲノム（genome）と呼ぶ。

遺伝情報を担うものは化学物質のDNA（デオキシリボ核酸）であり、DNAは4種類の塩基A、T、G、Cの一次元的配列で記述できる。ヒトの全染色体中にはDNA塩基配列が30億対あるとされ、その部分集合（領域）である遺伝子（gene）は約10万と言われ、各染色体上に密あるいは疎に散らばっている。

これらのヒト遺伝情報を全て解明しようという計画がヒトゲノム解析計画で、国際的な協調と競争の元に進められている。

まず、遺伝子あるいはそれらしきものの染色体上の位置を決める遺伝子地図作り（mapping）が行なわれ、これと並行あるいは前後して、その遺伝子の詳細塩基配列の決定（sequencing）が行なわれる。そしてその働き、機能が解明される。DNAの配列は蛋白質のアミノ酸の配列を決めている。

遺伝子の機能解明の中で遺伝病やがんなどの病気に対する原因究明、治療法の発見なども広く研究されている。

3. GDBの由来

遺伝子の地図情報のデータベース化は米国エール大学のHGML（Human Gene Mapping Library）に端を発し、その後、ハワード・ヒューズ医学研究所の支援の下に、ジョンズ・ホプキンス大学（JHU）でGDBとして開発・運用されている。GDBのオンライン提供の最初、Version1.0は1990年9月であり、現在改良が加えられVersion5.3となっている。

GDBは資金的には、米国DOE（エネルギー省）とNIH（国立衛生研究所）の援助を受けている。また、国際的援助も受け入れ、平成4年度から日本も科学技術庁から拠出金を出している。

ヒトゲノム解析においては、国際的ヒトゲノム機構（HUGO、Human Genome Organization）が重要な役割を演じている。遺伝子地図の作成では、染色体毎のワークショップ（SCW、Single Chromosome Workshop）と染色体調整会議（CCM、Chromosome Coordinate Meeting）を開催している。さらに、染色体毎に国際的なeditorを数名置きデータのGDBへの入力・評価を行なっている。

最近では論文発表の事前にGDBのId番号の取得を求められている。

4. GDBのデータ構造

情報提供の側面から見たとき、GDBは2つのデータベースから成り立っている。狭義のGDBとOMIMである。GDBの中心は遺伝子地図情報であるが、内容は多様で11のデータ・マネージャーと称するデータ区分でオンライン提供されている（図1）。中心となる情報は、遺伝子等の染色体上の位置情報の遺伝子座（Locus）マネージャー、遺伝子座間の位置関係及び距離を表す遺伝子座のセット情報の遺伝子地図（Map）マネージャーで検索・表示する。個々の遺伝子座の正常及び異常な状態における個体差、民族差は、多型（Polymorphism）、変異（Mutation）マネージャーで参照し、検定集団（Population）マネージャーでは検査を行なった民族等の集団を指定する。これらのデータを得るために用いられたDNAの断片、DNAクローンのセット、細胞などの研究材料は、それぞれプローブ（Probe）、ライブラリー（Library）、細胞株（Cell Line）のマネージャーに実験条件等も含めて収録され、連絡先（Contact）マネージャーで材料情報の提供者を知ることができる。各情報は、情報源である引用文献（Citation）及び、全データに付与されているId番号を管理するGDB Idマネージャーとリンクしている。

その中で、遺伝子座情報、遺伝子地図情報、引用文献情報について図２～４で例を示す。

また、GDBに格納されているデータの件数を表１、表２に示す。

OMIMは、遺伝病にかかわるVictor A. McKusick博士の著書、Mendelian Inheritance in Manのフルテキストデータベースでオンライン版である（図５）。

5. GDBの提供

GDBは、UNIX OSの元でリレーショナル・データベース管理システム（RDBMS）のSYBASEを用いて検索システムを構成している。提供用のサーバーとしては、SUN4/490あるいは690が用いられている。

平成６年２月に、JICSTの筑波のGDBセンターからGDBのオンライン提供が開始された。

利用方法としては、インターネット経由、ISDN経由、電話網経由が可能で、操作性の良いグラフィック端末のワークステーション、キャラクター端末としてMac、PCが使える。

研究支援ということで、国際的に無料で公開されており、日本でもそれに従っている。

6. 考察

(1) データの視覚化と操作性の向上

GDBでは、現在、文字列（character）の情報を扱っており、視覚、操作性の面で限界がある。GUI（Graphical User Interface）の活用により、ゲノム情報の視覚化（イメージ／画像利用）を図る要望が強く、各所で研究開発がなされつつある。

我々の「ヒト遺伝子地図作成技術の開発に関する研究」で作成されたシステムの例を、データ入力部（図６）、データ蓄積・表示部（図７）で示す。

(2) 関連データベースの統合検索

ヒトゲノム解析においては、幅広い領域において研究が行なわれているので、以下のような関連情報が必要とされている。

①シークエンス情報（DNA配列情報）
GenBank、EMBL、DDBJ、GSDB

②蛋白質アミノ酸配列情報
PIR、Swiss-plot

③蛋白質構造情報
PDB

④ヒト以外のゲノム情報
MGD、FlyBase、AceDB

⑤医学文献情報
MEDLINE

以上の中で、特に遺伝子地図情報とシークエンス情報の統合化が求められている。

しかし、統合化された単一データベースと、分散型データベースを検索の都度アクセスしに行く方法の長所・短所の比較が難しい。今後は、ネットワークの発達で分散型データベース指向の可能性が強い。しかし、多くのデータベースの仕様の変更をタイムリーにフォローするのも大変である。

(3) 日本の研究者からの要望

①日本語表記システム

利用者の層が広がると日本語表記のシステムが求められる（図７）。

②日本情報の補充

日本で発生した情報はMEDLINEでは、充分でないので補う必要がある。

GDBで収集されないデータの蓄積・管理を考える必要もある。

(4) ネットワーク活用の新しいデータベース提供技術

JHU等でGOPHER、WWW、WAIS等による方法が試みられている。

7. 謝辞

本稿は、科学技術振興調整費研究課題「ヒト遺伝子地図作成技術の開発」及び、理化学研究所「GDB（ゲノムデータベース）の開発・導入に関する研究」の共同研究成果による。

8. 参考文献

平川美夏他、GDB（ゲノムデータベース）—ヒトゲノム解析プロジェクトにおけるデータベースの構築—、情報管理　Vol.36, No.11, p.1023-1032（Feb. 1994）

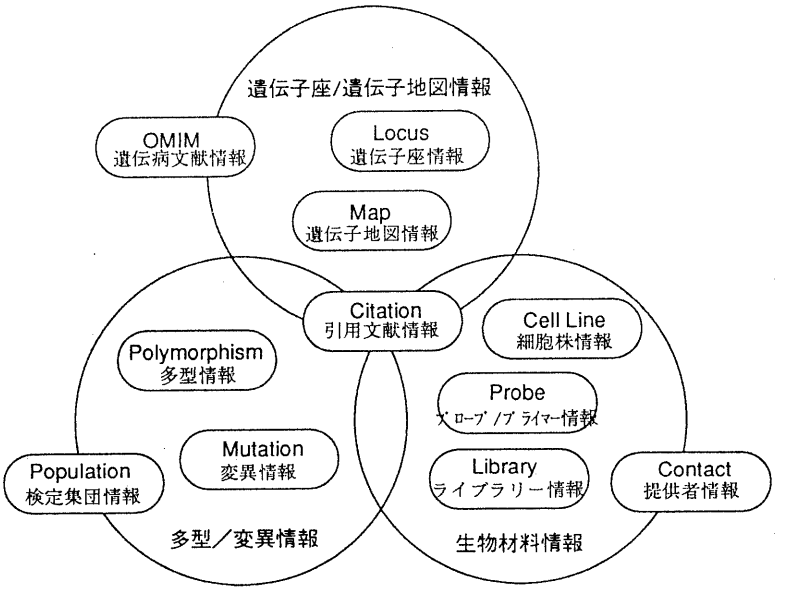


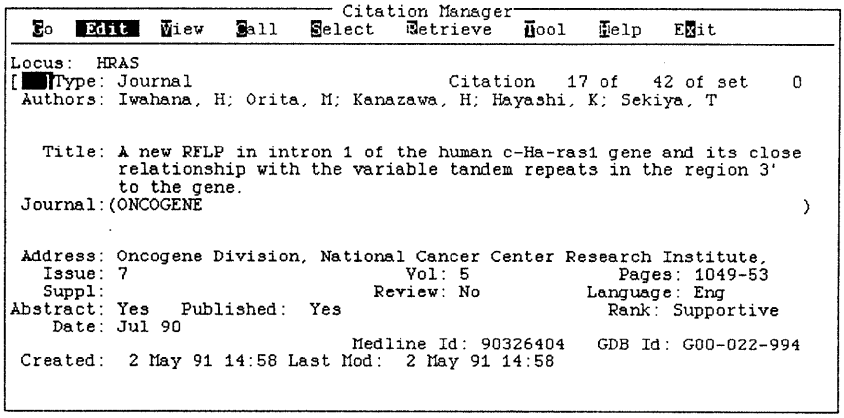
図1　GDBのデータマネージャー

```
                              Locus Manager
 Go   Edit   View   Mall   Select   Retrieve   Tool   Help   Exit

[ ]                                          Locus    1 of    6 of set    0
 Locus Symbol: HRAS                Previous HGM Symbol: HRAS1
   Locus Name: Harvey rat sarcoma viral (v-Ha-ras) oncogene homolog
Cyto Location: 11p15.5

      MIM #    Disorder
      190020

 Assign. Mode: (A,D,L,R,S          )          EC Number:
      GenBank: J00277,K00654,M19990             Probes: Yes
  Polymorphic: Yes      Het:        0.84     Allele Set: Bb
                                          Ref. Marker: Yes
    Citations: Nature 300:773-4 1982
               Science 219:498-501 1983
   Annotation:
      Created:  1 Jan 86 00:00          Last Modified:  9 Jun 93 15:27
                                                        GDB Id: G00-120-684
```

図2　遺伝子座(Locus)マネージャーのデータ画面

```
                               Map Manager
 Go   Edit   View   Mall   Select   Retrieve   Tool   Help   Exit
Locus:  HBB
[ ]                                          Map    1 of    7 of set    0

  Map Symbol: C11M1          Type: (Chromosome order )  No. of Elements:  22
      Method: (Reference                                              )
    Location: 11p15.5-qter
         Map: HRAS_INS-D11S454-D11S12_HBB_PTH-D11S455-D11S16-D11S417-D11S9-D
              11S436-D11S427-D11S534-D11S533-D11S388_D11S35-D11S424-DRD2_APO
              C3-D11S490-CD3D-D11S485



    Citation: Cytogenet Cell Genet :,1991
  Annotation:

     Created: 22 Aug 91 16:22           Last Modified: 23 Aug 91 11:22
Release Date:                                          GDB Id: G00-128-499
```

図3　遺伝子地図(Map)マネージャーのデータ画面



```
                              Citation Manager
 Go  Edit  View  Call  Select  Retrieve  Tool  Help  Exit
Locus:  HRAS
[ ]Type: Journal                          Citation   17 of   42 of set    0
 Authors: Iwahana, H; Orita, M; Kanazawa, H; Hayashi, K; Sekiya, T

   Title: A new RFLP in intron 1 of the human c-Ha-ras1 gene and its close
          relationship with the variable tandem repeats in the region 3'
          to the gene.
 Journal:(ONCOGENE                                                        )

 Address: Oncogene Division, National Cancer Center Research Institute,
   Issue: 7                          Vol: 5                 Pages: 1049-53
   Suppl:                          Review: No            Language: Eng
Abstract: Yes   Published:  Yes                            Rank: Supportive
    Date: Jul 90
                                 Medline Id: 90326404   GDB Id: G00-022-994
 Created:  2 May 91 14:58 Last Mod:  2 May 91 14:58
```

図4　　引用文献(Citation)マネージャーのデータ画面

表1　GDBのデータ別統計

Genome Data Base Data Statistics
Date: Apr. 15, 1994

Object	Totals
LOCUS:	
Total Genes:	4547
Total D-segments:	37405
Mapped Genes:	3219
Mapped D-segments:	28649
Mapped Fragile Sites:	116
Mapped Breakpoints:	463
Map Sets:	768
Total Mapped Loci:	33215
Total Loci:	43323
PROBE:	
PCR:	11849
ASO:	464
Clones:	79930
Total Probes:	92243
POLYMORPHISMS:	
Polymorphic Genes:	889
Polymorphic D-segments:	6737
Total Polymorphisms:	10609
CITATIONS:	
Journal Articles:	33073
Personal Communications:	6698
Abstracts:	822
Books:	42
Theses:	3
Total Sources:	40638
Total Sources linked:	22546

表2　GDBの染色体別統計

DATE: Apr. 15, 1994

The total number of loci includes fragile sites, breakpoints, and maps.
Therefore, on some chromosomes this number is greater than the sum of the
genes and D-segments.

Chromosome	Genes	D_segments	Total
Unassigned	1328	8756	10084
1	342	1188	1572
2	181	1277	1533
3	127	2312	2518
4	129	2308	2521
5	116	1438	1585
6	175	946	1180
7	145	1513	1732
8	96	1032	1164
9	133	774	958
10	104	1147	1313
11	218	1897	2180
12	175	831	1025
13	51	640	734
14	94	545	656
15	83	530	633
16	119	681	881
17	175	1141	1370
18	34	802	884
19	197	549	762
20	65	451	541
21	42	1413	1516
22	95	548	672
X	254	4182	4688
Y	23	583	630
MT	55	0	55

```
DOCUMENT READER:                                              OMIM (1, 08Oct93)
  F(screen Forward), B(screen Backward), J(Jump to next section),
  S(Show query terms), =(Search words), M(gene Map), Z(defects list),
  L(back to selection List), Q(new Question), P(Print/output), ??(help), E(Exit)
--------------------------------------------------------------------------------

-- Document 1 of 1 (7712 lines) --
    *141900 HEMOGLOBIN--BETA LOCUS [HBB; SICKLE CELL ANEMIA, INCLUDED;
    BETA-THALASSEMIAS, INCLUDED; HEINZ BODY ANEMIAS, BETA-GLOBIN TYPE,
    INCLUDED; METHEMOGLOBINEMIA, BETA-GLOBIN TYPE, INCLUDED; ERYTHREMIA,
    BETA-GLOBIN TYPE, INCLUDED; DYSERYTHROPOIETIC ANEMIA, CONGENITAL,
    IRISH OR WEATHERALL TYPE, INCLUDED]
--------------------------------------------------------------------------------

    The alpha and beta loci determine the structure of the 2 types of
    polypeptide chains in adult hemoglobin, Hb A.

    By autoradiography using heavy-labeled hemoglobin-specific messenger
    RNA, Price et al. (1972) found labeling of a chromosome 2 and a group
    B chromosome.  They concluded, incorrectly as it turned out, that the
    beta-gamma-delta linkage group was on a group B chromosome since the
    zone of labeling was longer on that chromosome than on chromosome 2
    (which by this reasoning was presumed to carry the alpha locus or
    loci).  Study of a case of the Wolf-Hirschhorn syndrome (4p-)
    suggested that the B group chromosome involved is no. 4.  Barbosa
```

図5　　OMIMのデータ画面

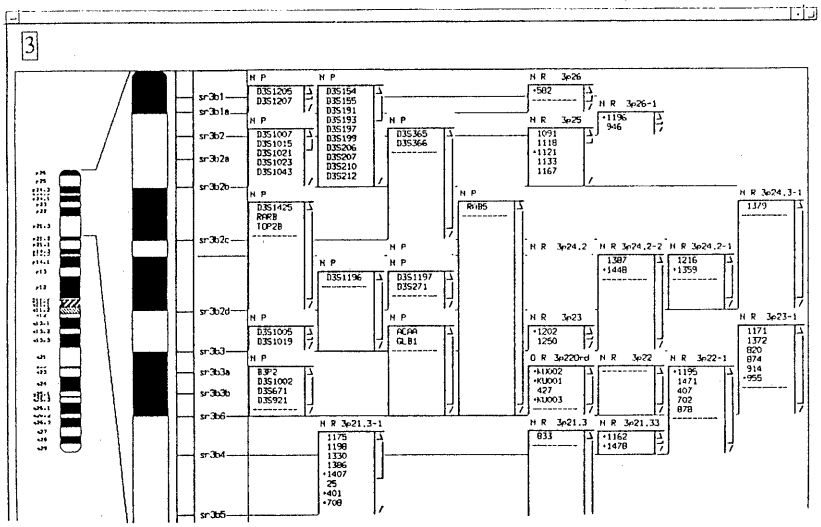


図6　　入力部の画面例

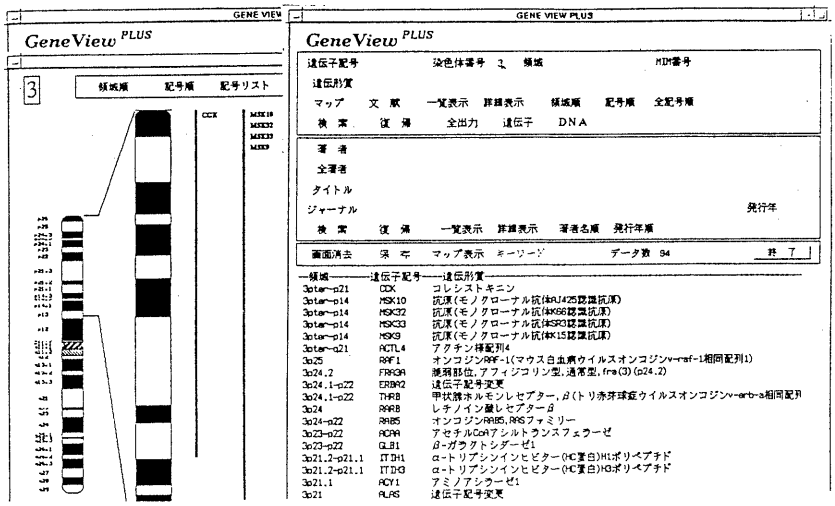


図7　蓄積・表示部の画面例