データベース・システム 84－6
（１９９１．７．１７）

ゲノム情報のデータベース化に向けて

高木利久* 佐藤賢二* 久原哲† 古川哲也‡

*九州大学情報処理教育センター
†九州大学大学院農学研究科
‡九州大学大型計算機センター

概要

近年、各種生物のゲノム(genome: 個々の生物の全遺伝情報を担う染色体)を実験的に解析することが可能になってきた。各国におけるゲノム解析プロジェクトの推進により、ゲノムに関する情報は今後爆発的に増加することが予想され、これに伴い、これらのデータをデータベース化したいという要求が高まりつつある。しかしながら、ゲノム情報は、従来のデータベースで扱われてきた事務データやエンジニアリングデータとは異なる性質を少なからずもっており、従来のデータベース技術だけで対処することは困難である。本稿では、ゲノム情報の性質を述べ、それらのデータベース化に際しての問題点を明らかにする。

Towards Development of Database Systems for Genomic Information

Toshihisa TAKAGI*, Kenji SATOU*, Satoru KUHARA† and Tetsuya FURUKAWA‡

*Educational Center for Information Processing, Kyushu University
e-mail: {takagi, satou}@ec.kyushu-u.ac.jp
† Graduate School of Genetic Resources Technology, Kyushu University
e-mail: kuhara@grt.kyushu-u.ac.jp
‡Computer Center, Kyushu University
e-mail: furukawa@cc.kyushu-u.ac.jp
6-10-1, Hakozaki, Higashiku, Fukuoka, 812, Japan

Abstract

Genome is a set of choromosomes which is a molecular blueprint for the building of an individual organism. With the progress of study on the genome, the role of database systems for genome analysis has been of great importance. However, genomic information has some features different from those of business and engineering data which can be handled by conventional database technologies. In this paper, we describe the features of the genomic information and point out some problems that confront us at the time of the design and implementaion of genome databases.

1. まえがき

組換えDNA技術を始めとするバイオテクノロジーのめざましい発展により、各種生物のゲノム(genome: 個々の生物の全遺伝情報を担う染色体)を実験的に解析することが、近年可能になってきた。このような状況の中で、いくつかの生物のゲノムを対象にして、その全構造を解析し、すべての遺伝情報を解読しようとする試みが各国で活発に進められている[Mat91]。日本では、平成3年度より創成的基礎研究(代表者: 松原謙一・大阪大学教授)および重点領域研究(代表者: 金久實・京都大学教授)を中心とした文部省ヒトゲノムプロジェクトが5ヶ年計画で始まった。また、農水省ではイネのゲノム解析計画が進められている。この他、科学技術庁、厚生省、通産省においてもこれらに類した計画が進行中である。一方、米国や欧州では、日本に先駆けてゲノム解析プロジェクトが発足しており、莫大な費用を投入して、データの収集や解析用ソフトウェアの開発が行われている。

人間の生物学的な設計図とも呼べるヒトゲノム(Human genome)は、22本の常染色体とX,Yの性染色体とからなり、これには人間のすべての遺伝情報が書かれている。ゲノムの本体はDNAであり、DNAは基本的には4種類(A,T,C,G)の塩基によって構成される。ヒトゲノムの場合、その本体は約30億対の塩基からなる巨大なDNA分子であり、その中には5万から20万の遺伝子(gene)が配列されていると推定されている。一つの遺伝子は、1千から1万個程度の塩基から構成される[Lew90]。

DNA分子は、ワトソンとクリックの研究でよく知られているように、二重らせんと呼ばれる3次元の構造をとっている。これは、縄ばしごをねじったような構造であり、その縄に相当するのが、塩基の並びである。4種類の塩基のうち、どの塩基とどの塩基とが対になるかは一意に決まっているため、つまり、片方の縄における塩基配列が決まれば、もう片方の縄の塩基配列も決まるため、DNA分子は、1本の縄、すなわち、1次元の塩基配列(文字列)によって表現できることになる。

この1次元文字列上には、生物が生命活動を営む上で必須の分子(例えば、タンパク質)に関する情報と、それらがいつどのような状況の下で発現するべきかといった制御情報などが書かれている。しかしながら、DNAの1次元文字列だけからこれらの高次の情報を解読する方法は、タンパク質をコードする規則を除いて、ほとんど分かっていないのが現状である。現在最もよく行われている解析方法は、類似配列による推定である。これは、同じ、あるいは、異なる生物種間で、DNAやタンパク質の1次元配列を比較し、その中から類似した配列を捜し(ホモロジー検索と呼ぶ)、これにより既知のものから未知のものを推定する方法である。しかしながら、ホモロジー検索で検出されるのはあくまでも1次構造の類似関係であり、それらの生物学的な意味については、分子生物学者がその他の知識や実験データをもとに判断しなければならない。

このように、ゲノム解析においては、DNAの1次元配列データだけでなく、実験や遺伝学によって得られた遺伝子地図や物理的地図、タンパク質に関するデータなどを総合的に参照しながら研究を進める必要がある。ゲノムの解析に関わるこれらの情報を本稿では総称してゲノム情報と呼ぶことにする。

ゲノム解析プロジェクトは米国を中心として既に始まっており、この中でゲノム情報に関する各種のデータが現在急速な勢いで蓄えられつつある。例えば、GenBank(米国ロスアラモス研究所)と呼ばれるデータバンクには、DNAの塩基配列に関するデータが蓄えられている。また、PDB[BKWMBRKST77](米国ブルックヘブン研究所)はタンパク質の3次元構造の、GDB(米国ハワードヒューズ医学研究所)は遺伝子地図の、データバンクである。これらのデータは、通常、テキスト形式で蓄えられている。(DBMSによって構造化・組織化されたデータベースと区別するため、テキスト形式で蓄えれているデータの集まりを本稿ではデータバンクと呼ぶことにする。)

各国におけるゲノム解析プロジェクトの推進により、ゲノム情報は今後爆発的に増加することが予想されている。そのため、これらのテキスト形式のデータをデータベース化したいという要求が高まりつつある。現に、GenBankは、関係DBMSのSybaseを用いてデータベース化が行われている。しかしながら、ゲノム情報は、従来のデータベースで

扱われてきた事務データやエンジニアリングデータとは異なる性質を少なからずもっている。また、GenBankやPDBをそれぞれ独立にデータベース化するだけでは不十分である。なぜなら、これらのデータバンクに格納されているデータは、相互に複雑で密接な関連をもっているからである。これらの問題に、従来のデータベース技術だけで対処することは困難である。

本研究の目的は、遺伝情報の解明を支援するためのデータベースを開発することにある。このようなデータベースを開発するには、新しい技術を必要とする。そのため、本研究は、分子生物学にとってだけでなく、データベース研究の観点からも意義深いものと考えられる。

本稿では、ゲノム情報のデータベース化に向けて、その問題点を明らかにすることを目指す。まず、2節では、分子生物学の観点から、ゲノム情報とは何か、それらの情報がどのように関連しているかを、各国におけるデータ収集状況を交えて述べる。次に、3節では、ゲノム情報のデータベース化に関する従来の研究を紹介する。4節では、従来のデータベースが扱ってきた事務データやエンジニアリングデータと比較しながら、ゲノム情報の性質を述べ、それらのデータのデータベース化に際して何が問題となるかを検討する。最後に、5節において今後の課題をまとめる。

2. ゲノム情報

2.1 ゲノム情報とは

生命現象や遺伝のメカニズムの解明は、人類の大きな夢であり、いままで多くの研究が行われてきた。その結果、上は系統発生や種のレベルから、下はDNAの塩基配列や遺伝子の発現調整のレベルまで、多くのことが分かってきた。しかしながら、データや技術の不足から、まだ、遺伝情報を完全に解読するには至っていない。

生物の遺伝情報はすべてDNAにコード化されていると考えられている。そのDNAは、基本的には4種類の塩基(文字)の配列で表現できる。そのため、遺伝情報のすべてを解読するには、まず、対象とする生物の塩基配列のすべてを決定しなければならない。しかしながら、塩基配列がすべて分かっている高等生物は、いまのところない。例えば、ヒトゲノムは約30億の塩基から構成されるが、現在までに配列が分かっているのは、その中の0.5パーセントにも満たない。今後、この配列すべてをバイオテクノロジーを用いて決定する必要がある。配列すべてを決定することは、各国におけるゲノム解析プロジェクトの目標の一つである。

一方、これと並行して、塩基配列の、どの部分に、どのように情報がコード化されているか、を明らかにしなければならない。ヒトゲノムの場合、約30億の配列のうち、遺伝子に相当する部分は5パーセント程度しかなく、大部分は無意味な配列だと考えられている。遺伝子が配列中のどこに書かれているかを、配列から決定する方法はまだ分かっていない。また、遺伝子をコードする配列自体をとっても、その中にはプロモータ(読み始めを指示する部位)、エクソン(タンパク質のアミノ酸配列を指示する部位)、イントロン(意味のない部位)などの部分配列(これらを機能部位と呼ぶ)があるが、これらを構成する規則も明らかにはなっていない。

塩基配列は、4種類という非常に少ない文字種で記述されているため、単に文字列を眺めるだけではこの問題は解決しない。いままでの実験や研究で得られた各種のデータや知識を総動員して、解読する必要がある。塩基配列の解読に関与するデータで、主なものを以下に示す。

- 生物学的情報: 生物学の分類データなど。
- 家系図: 遺伝病などをもつ家系のデータ。
- 遺伝子地図: 遺伝子間の論理的距離に関する地図。
- 制限酵素地図: 制限酵素によるDNA配列の物理的位置の地図。
- クローン情報: DNA断片試料に関するデータ。
- DNA(RNA)の構造: 1次構造(塩基配列)、2次構造、3次構造、機能部位。
- タンパク質の構造: 1次構造(アミノ酸配列)、2次構造、3次構造。
- 文献情報: 各種のデータを掲載した文献に関するデータ。

これらのデータの関連の一部を図1に示す。図1に示すようにこれらのデータは密接に関連している。

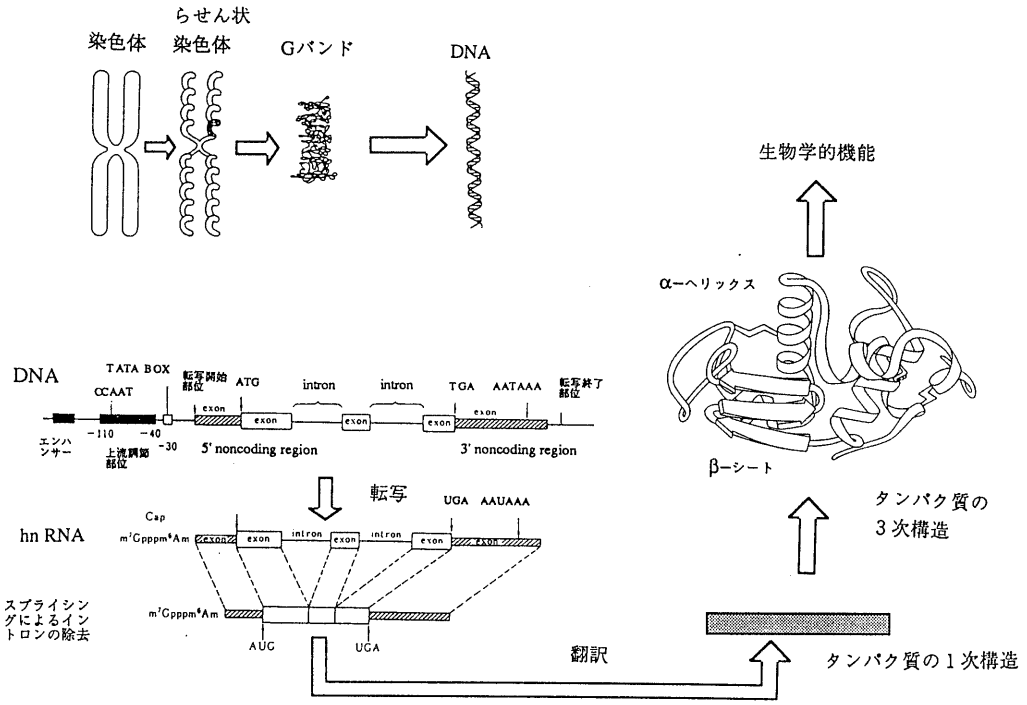

図 1: ゲノム情報の関連図

2.2 ゲノム情報の収集状況

表1に、ゲノム情報に関するデータバンクの内で主要なものとその収集機関、収集データの種類を示す。例えば、GenBankは、米国のロスアラモス研究所が中心になり、分子生物学の雑誌あるいは直接研究者から、塩基配列データとそれに関連する情報を収集し、配布している。これらのデータバンクは、収集しているデータの種類によっていくつかのグループに分けられるが、いずれも機械可読な情報をテキストファイルの形で提供するという点では共通している。同じ種類のデータを収集しているデータバンクの中には、参照すべき雑誌の範囲を分けるなどして、互いに協力しているところもある。例えば、GenBank, EMBL および DDBJ は、塩基配列データの共有や配布に関して協力体制をとっている。そのため、研究者はいずれかのデータバンクに登録すればよいことになっている。

しかしながら、同じ種類のデータを集めているにもかかわらず、データバンクによって、データのフォーマットや内容が微妙に異なり、統一がとれていないこともある。フォーマットの統一の問題に関しては、入力データの記述にBNF記法を使う方式も提案されている[MEG89]。例えば、EMBL/GenBank共通属性テーブルの定義がそれに当たる。

1970年代にDNAの塩基配列決定技術が開発され普及して以来、DNAの配列データの量は増加の一途を辿っている。例えば、GenBankが蓄積しているデータの量は指数関数的に増大している。今後、塩基配列決定の自動化技術が進めば、データの増加率は、ますます大きくなることが予想される。

3. データベース化に関する従来の研究

2.2節で述べたように、データバンクには基本的にテキスト形式でデータが格納されている。大部分の研究者は、これらのデータをそのままの形で譲り受け、各自が用意したプログラムを使ってその中から必要な情報を抽出しているのが現状である。しかしながら、データベース化を試みた例もある。以下に、それらの研究事例をいくつか紹介する。

名称	収集機関	種類
GenBank	ロスアラモス研究所	DNA1次
EMBL	ヨーロッパ分子生物学研究所	DNA1次
DDBJ	国立遺伝学研究所(日本)	DNA1次
PIR	米国基礎医学研究財団	タンパク1次
SWISS-PROT	ジュネーブ大学	タンパク1次
JIPID	東京理科大学	タンパク1次
PDB	ブルックヘブン研究所	タンパク3次
GDB/OMIM	ハワードヒューズ医学研究所	遺伝子地図
ATCC	全米系統保存施設	試料情報
	コリエル医学研究所(米国)	試料情報

表 1: ゲノム情報のデータバンク

GENAS[KMFFSTS84]

GENAS は GenBank, EMBL, PIR および PDB をデータ源とする、DNA とタンパク質とに関する推論型の関係データベースシステムである。GENAS は、以下の 4 つの部分からなる。

i) 抽出した塩基配列、タンパク質アミノ酸配列、立体構造データおよびそれらに関する文献情報などを格納したデータベース
ii) パッケージ化された 26 個の応用プログラム
iii) ホーン集合に基づく推論機構
iv) 文献検索システム型インタフェース

例えば、塩基配列の検索を行う場合、利用者は、文献検索システムで良く使われる FIND 等のコマンドか、または、ホーン節に基づく論理型言語で質問を行い、DNA 塩基配列の集合を得る。得られた集合に対して応用プログラムを実行することにより、塩基配列の比較や制限酵素切断部位の検索、あるいは 2 次構造の予測などが行えるようになっている。GENAS は、現在、九州大学大型計算機センターの FACOM M-780 上で稼働している。

BIPED[IsS89, StI89]

BIPED は PDB をデータ源とするタンパク質の構造に関する関係データベースシステムである。Micro Vax II 上に実現された BIPED システムは、DBMS として ORACLE を使っている。質問は SQL で記述する。ORACLE で管理する関係テーブルは 10 個で、それぞれが 5 ないし 183 個の属性を持つ。294 個の PDB テキストファイル(50MB) に対するデータベースの大きさは、テーブルが 240MB、索引ファイルが 33.1MB、格納されるタプルの数は 1,027,360 個である。

Morffew らの研究 [MTS83, MoT86]

彼らは、タンパク質の構造データの格納・検索に初めて関係データベースを用いた。そのシステム PRTV は、グラフィックシステムと合わせて利用することが可能で、ユーザは質問を関係代数の形で入力し、検索結果をタプル集合またはグラフィックの形で得ることができる。データ源は PDB である。

その後、彼らは、タンパク質の構造データの検索に Prolog を用いた。このシステムでは、8 種類の述語を使ってタンパク質の構造データを記述する。利用者がルールと質問を与えると、システムは質問の内容に従って PRTV を検索し、その結果をファクトに変換し、その後質問処理を行う。このシステムでは、主記憶容量の制限により、一度に 1 種類のタンパク質に関するファクトだけを読み込む。

Gray らの研究 [GPKF89]

彼らは、タンパク質の構造データのために、オブジェクト指向データベースを用いた。このシステムは、関数型データモデルを採用したデータベースと Prolog とから構成される。質問は、関数型質問言語 Daplex と Prolog との両方で行うことが可能となっている。オブジェクトは、タンパク質の構造データ間にある階層を反映するように設計されている。例えば、helix と呼ばれるタンパク質 2 次構造には、α-helix や π-helix などの種類がある。彼らは、このような階層性をオブジェクト間の関係で表現することを試みている。なお、データ源としては、BIPED プロジェクトにおいて作成されたものを用いている。

Rawlings らの研究 [RTNFS86, Raw89]

彼らは、タンパク質の構造、とくに β 鎖を含む超 2 次構造の検索に Prolog データベースを用いた。PDB から FORTRAN プログラムを使って、α 炭素原子の 3 次元座標、アミノ酸配列、2 次構造配列内での位置、β 鎖の隣接関係や方向(並行／逆並行) などのデータを抽出し、これらをファクトとして格納

する。また、タンパク質の構造に関する知識をルールとして記述する。例えば、「アミノ酸配列の上で連続していて、かつ相対的な位置関係では隣接していて、かつ逆並行の向きにある2本のβ鎖は、超2次構造βヘアピンを構成する」とか、「連続する2個のβヘアピンは、β屈曲構造を構成する」という知識をルール化している。これを使って超2次構造の検索を行う。また、このシステムにはグラフィックインタフェースがあり、検索したい超2次構造の編集や検索結果の表示などがマウス操作で行える。

PACADE[KSFTTS91]

PACADEはPDBをデータ源とするタンパク質立体構造のための演繹データベースシステムである。このシステムは、ボトムアップ評価に基づく推論エンジンと関係データベースとから構成されている。関係DBMSとしてはSybaseを用いている。推論エンジンはファクトの形式に変換されたデータベース検索結果に対して推論を適用し質問処理を行う。推論エンジンは、マジックセット法に基づくルール変換機構をもっており、これにより推論中に生成される不要な中間ファクトの数を減らし、実行効率をあげている。

4. ゲノム情報の性質とデータベース化の問題点

以下では、従来のデータベースが扱ってきた事務データやエンジニアリングデータと比較しながら、ゲノム情報の性質を述べるとともに、それらのデータベース化を図る上での問題点を探る。

(1) 種類の多様性

2節で述べたように、ゲノムの解析を行うには、多種多様なデータを扱う必要がある。その一例として、図2にGenBankの1エントリのデータを示す。この中には、DNAの1次元配列データ、その機能部位に関するデータ、それを掲載した文献に関する情報などが含まれている。塩基配列を扱うにはテキストデータベースの手法が、文献情報を扱うには文献検索の手法が、機能部位情報を扱うには関係データベースの手法が、向いていると思われる。効率のよいデータベースを実現するには、これらの異なる手法をどのように統合するかが問題となる。さらに、GenBankには含まれないが、ゲノム解析では、以下のデータも扱う必要がある。

- 図形データ(遺伝子地図、物理地図)
- 座標データ(DNAやタンパク質の3次構造)
- 階層構造データ(生物学の分類情報やタンパク質の構造)
- 画像データ(各種の実験データ)

```
LOCUS       ACCP322P    1337 bp ds-DNA          SYN      15-SEP-1990
DEFINITION  Synthetic plasmid pWH1266 origin of replication (ori) region.
ACCESSION   M36473
KEYWORDS    .
SOURCE      A.calcoaceticus lwoffi plasmid and pBR322 DNA, clone pWH1266.
  ORGANISM  Cloning vector
            Artificial sequences; Cloning vehicles.
REFERENCE   1  (bases 1 to 1337)
  AUTHORS   Hunger,M., Schmucker,R., Kishan,V. and Hillen,W.
  TITLE     Analysis and nucleotide sequence of an origin of DNA replication in
            Acinetobacter calcoaceticus and its use for Escherichia coli
            shuttle plasmids
  JOURNAL   Gene 87, 45-51 (1990)
  STANDARD  simple staff entry
FEATURES             Location/Qualifiers
     rep_origin      310..337
                     /note="origin of replication"
BASE COUNT     447 a    229 c    251 g    410 t
ORIGIN

   1  gatcgtagaa atatctatga ttatctgaa  gaacgcaacc ctatagcagc tattgaaatt
  61  gatgatttaa ttgaagaaaa gacagattta gttgttgata atcgactgat ggggcgcaca
 121  ggcagacaga aagatactag ggagttagtg atacatccgc attatgtggt tgtatatgac
 181  atcactgata taatacggat actcagagtg ctacacacat cgcaggagtg gtcatgactt
 241  actcatgtac ttggattat ttagtgttat aaaatcctga ttiatiaaatt ttttttgtta
 301  aaaaagataa aagcccctg caattgctg gggcttiacc gtaatttatg ggglacagat
 361  ctlcgatact gacatatcgg caatcgaaag cattaaggtt tgacgacegc taatgattic
 421  accacagggg cttaatgtac ctgtcttaaa ttctaaggtt ttaactcgct ttgtcaagca
 481  tagaccccaa aaatttagcc aatgtctgta actcaatctg tccatgtgtg ggtgatgagg
 541  tacagtgacg ctagcacaca tcggaaaaac gctattacta ggggaacgta acagagtagc
 601  ggacgcaatg agtagtcatt taattggcgg ttatgagcgt gttcaggcgg tgctatcaat
 661  cgtaatcata acagtggcag cttgatacag tgatgtcatc cctgatgcga aagcgaccga
 721  ccgacggtac atcgaatggg aatacttag ggtgatttti aagaatcgct ctagggtgag
 781  tatttcccat tcagctctgc tcccticcic tggtactita atcaaaagca ctactaaaca
 841  tatgttttta aataaaaaat attgatatag agataatatt agtaagaata attaacaat
 901  tgaatataga taaatcattg ttaaataaag attaattatt aaaatgaatg tatacttata
 961  tataaatcaa tgattiaaaa tattgataa agaaaacttt tcaaaaaaaa tataattgag
1021  attgtgtcat ttcggtcaat tcttaatatg ttccacgcaa gtttagcta tggtgctaaa
1081  cagaaatttg ctgaaaaaga actttcact gaactggtta aaatgtaagc agcctgagag
1141  ccgccaaaaa ttttaaaaac aaaccgcctt aatcatcttc aaaaaatacc tctaaaacct
1201  caccattigc gtttaagac ccatatttca tcctgcccti atgttcccat gctgatagct
1261  ataaagtgtc tgtaatcgct tcctatgacg ttctaggctg ttgataacti ttggaacaac
1321  gcaaaatgtt aaaatcc
```

図2: GenBank データの例

(2) 重複

塩基配列のデータを始めとして、ゲノム情報は個々の研究者が自分の興味のある箇所から解析・収集を行っており、系統的、組織的にそれらが行われているわけではない。そのため、同じ遺伝子に対して、いくつもの解析結果が得られている場合がある。これにより、異なるデータバンク間にも、同じデータバンク内にも、データの重なりがある可能性があり、重なり具合いもまちまちである。これらのデータの一部は整理・統合されているが、そうでない場合も少なくない。このようなデータの整合性を

どのようにとればよいかは問題である。重複したデータを排除することの困難さは、それが実験対象の個体差に由来するものかどうか判断できないことによる。これについては、以下の(5)で述べる。

(3) 粒度

図2の例からは、分からないが、GenBankの1エントリに含まれる塩基配列は、文字数にして、短いものは数十、長いものは数十万の長さがある。このようにデータの粒度が大きく異なるデータをどのように効率よく扱うかは、データベース化を図る上で大きな問題である。

(4) 複雑な関連

2節で述べたように、塩基配列をただ眺めただけでは、遺伝情報を解読することはいまのところできない。これは、異なるデータバンク(データベース)に格納されている情報も参照しなければならないことを意味する。例えば、塩基配列のエクソンやイントロンの構成に関する規則を見つけるには、タンパク質のデータを参照する必要がある。これらのデータベース間にまたがるリンクをどのように管理するかが問題である。しかも、2.2節で述べたように、データの標準化が図られていない現状では、このことはより一層困難を伴う。また、同じデータバンク内のデータも密接に関連していることがある。例えば、GenBankでは、ある遺伝子が複数のエントリに分割されて格納されていることがある。そのため、その遺伝子に着目して処理を行いたい場合は、遺伝子ごとにデータを繋合わせることが必要となる[NSSUS91]。このような処理のためにデータベースをどう構成しておくかも大きな問題である。

(5) 不確実性および個体差

ゲノム情報の中には、従来のデータベースで扱ってきたデータにはあまりない、不確実性が含まれる。例えば、DNAの塩基配列データは、実験によって求められるが、実験の精度からくる誤り(0.5パーセント程度)を常に含んでいる可能性がある。また、実験データをデータバンクに登録する際にも転記ミスが入り込む可能性がある。後者のミスはフロッピーディスクやネットワークを介することによってほとんど防ぐことができる。しかし、文字種が少ないことや文字列を構成する規則が分かっていない現状では、一旦ミスが混入すると検出するのは非常に困難である。また、DNA配列には当然個体差がある。DNAの解析にどの個体の染色体を用いるかによって、実験結果が異なる。同じ遺伝子を解析しているにもかかわらず、研究者によって得られたデータが異なる場合がある。この場合、どちらかが実験でミスを犯したのか、それとも個体差なのかは判断することはできない。そのため、どちらが正しいとも言えず、データバンクには、二つの実験結果を併記せざるをえない。このようなデータをどのようにデータベースに格納すべきか、また、どのように処理すべきかは、重大な問題である。

(6) データ量と今後の増加傾向

名称	データの種類	データ量
GenBank	塩基配列	エントリ数 43,903 総塩基数 55,169,276
PIR	タンパク質アミノ酸配列	エントリ数 28,232 総残基数 8,076,497
PDB	タンパク質立体構造	エントリ数 662
GDB	遺伝子地図	遺伝子数 1,878 DNA断片数 4,944

表2: 主要なデータバンクのデータ量

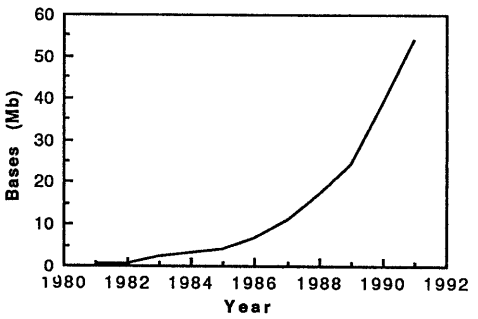

図3: GenBankのデータ量の増加傾向

表2に主要なデータバンクに登録されているデータ量を、図3にそれらのデータの増加傾向を示す。これらのデータ量は、従来の大規模データベースにおけるデータ量と比較するとそれほど大きいとは言えないが、データの種類の多様性やデータが複雑に関連していること、その上で複雑な検索や推論を必要とすることを勘案すると、少ないとは言えない

であろう。なお、データバンクの更新は、GenBankを例にとると毎日行われている。

(7) 処理の多様性

ゲノム情報処理の大きな特徴の一つに、データの操作や処理の多様性が挙げられる。例えば、塩基配列の検索一つをとっても、従来のデータベースで用いられてきた、完全一致による検索以外に

- キーワード検索: 通常の文献検索システムで行われているようなキーワードによる検索。
- モチーフ検索: モチーフと呼ばれる特定パターンを配列データから探す。文字列の置き換えは許すが空白の挿入・削除は許さない。
- 近似(高速)ホモロジー検索: 文字列の置き換え、および空白の挿入・削除を許す、類似配列の検索。ただし、厳密ホモロジー検索とは異なり完全性は要求されない。
- 厳密ホモロジー検索: 文字列の置き換え、および空白の挿入・削除を許す、類似配列の検索。

などが必要である。このような検索が必要なのは、上の(5)で述べたように、生物学的なデータは常に実験の誤り、個体差、種間差、無意味な配列などを含んでいるからである。類似している部分はどこか、あるいは、何パーセント似ているか、などが研究上の仮説を立てる際に重要になってくる。処理の多様性はこれらの検索方式の多様性だけに留まらない。ゲノム解析においては、研究者ごとにデータに対する処理が大きく異なるという特徴がある。これは、ゲノム用のデータベースが研究的な色彩が非常に強いことに由来する。データベースの特徴の一つは、データや処理の共有にある。各研究者の要求が、大きく異なる場合に、どの程度まで、DBMSがサポートすべきか、どの部分は応用プログラムに任せるか、は大きな課題である。

(8) 分散性

(4)で述べたように、ゲノム解析のデータは、複雑に関連しており、管理の面から言えば、どこか1箇所で集中管理することが望ましい。しかしながら、これらのデータベースを管理するには莫大な費用や人的資源を要すること、歴史的な経緯、各国や省庁間の障壁などを考慮すると、データバンクやデータベースは分散化せざるをえない。緊密なつながりをもつデータの分散化を如何に図るかが問題となる。

(9) 公共データと個人用データ

ゲノムの研究者は、GenBankなどの公共のデータ以外に、個人用データをもっていることが少なくない。この中には、その研究者自身が実験して得た未発表データもあれば、公共のデータを自分用に加工して作成したデータもある。個人用データをどのように管理するかは、個人の問題ともいえるが、ゲノム研究においては、このようなデータを簡単に扱える小回りのきくデータベースも必要である。また、個人用データの中には、秘密にしておきたい期間が過ぎれば、公共データベースに登録する可能性のあるものもある。公共のデータベースは個人用データが容易に取り込めるようになっていることが望ましい。

(10) 処理要求の不明確さ

効率がよく、使い易いデータベースを開発するには、データに対してどのような処理を行いたいか、あらかじめ明らかになっていなければならない。しかしながら、現時点ではゲノムの研究者でさえも、データに対してどのような処理を行えばよいのか明確ではないように思われる。そのため、研究が進むと、新たな種類のデータが発生したり、データベースの見方を根本的に変更したいという要求が起きたりする可能性がある。データベースは、このような場合にもある程度対処できるような構成になっていることが望ましい。

5. 今後の研究課題

(1) データのモデル化

ゲノム解析に必要な多様なデータをデータベース化する際に、どのようなデータモデルが適しているかを考えなければならない。4節で述べたように、値、文字列、図形、画像といったデータを扱うが、単に扱うデータがマルチメディアデータであるというだけではない。文字列や図形、画像の中で値がどの位置にあるかといった情報や図形と画像の対応関係など、メディア間の関連も必要となる。ま

た、データ長の不均一性の扱いやデータの重複の表現法も、要求されるデータ処理とも関連させて考えなければならない。

(2) データ処理

データベースを用いて行う処理は、研究者ごとに異なるものが要求される。これらの要求をデータベースのビューの概念ですべて吸収するのは困難である。もし複数の処理モデルが必要ならば、それらをどのように組み合わせるのかが問題となる。また、要求される処理も定まったものではなく、データを解析するための適当な(いわば思いつきの)操作を施すことも要求される。そのためには、データベースは、処理法が限られてくるようなものでなく、制約の少ない柔軟なものとする必要がある。

(3) 可変性への対処

ゲノム解析に関する研究は、いまだ途上にあり、データの性質やデータベースに対する要求が固まっているわけではない。データベースを構築した後も、その枠組みには納まらない様々なデータが発生することが予想される。そのような場合には、データベースの再構成の必要性も生じるが、それに対処できる方式を考える必要がある。データ個々についても、実験結果や実験結果からの予測によるものについては誤りが発見される、あるいは、データとデータとの統合が可能となる場合もあり、そのようなデータの削除・統合、およびそれによる他のデータへの影響を管理できるような方式を検討する必要がある。

(4) 推論機能

ゲノム情報を解析する上で、今後どのような推論機能が必要になるかはいまのところ明らかではない。しかしながら、従来研究されてきた推論機能では不十分であることは確かであろう。いままで研究されてきた推論機能(例えば、演繹データベース)は、基本的には、データの完全一致に基づいて、データ間の関係を調べるものである。しかしながら、生物学的データには、誤りや個体差がある。また、欠如しているデータも多い。このような状況の中でも、ある程度の推論が行えるような頑健な推論機構を開発する必要がある。

(5) スーパーコンピューティングとの連携

ホモロジー検索は大量の計算パワーを必要とする。例えば、あるDNA塩基配列とデータベース中のすべての塩基配列とを比較するとすると、スーパーコンピュータを用いても、近似的な方法でも1時間、厳密な方法では10時間程度の計算時間を必要とする。また、タンパク質の1次元アミノ酸配列から3次構造を予測するためには、スーパーコンピュータを用いて数時間が必要である。そのため、各種データベースとスーパーコンピュータとがうまく連携できるようにする必要がある。

(6) 分散化と統合化

4節で述べたように、ゲノム解析に関わるすべてのデータベースを一局管理することはできない。しかし、各種のデータは、密接に関わっている。これには統合化が必要である。分散化と統合化とをどのように実現するか大きな課題である。

(7) 知識の整備

高度なゲノム解析を行うには、実験データだけでなく、各研究者がもっている分子生物学に関する各種の知識も、機械可読な形式になっていることが望ましい。とくに、各研究者がもっているゲノム解析の戦略を共有できるようにすることが大切である。これらの知識をデータベースの検索にどのように活かせるかはともかく、今後、このような知識ベースを整備していく必要があろう。

(8)DBMSと応用プログラムとの境界

ゲノム解析用のデータベースは、研究的な色彩が強く、多数の研究者が共通して使用するデータやその操作は、明確ではない。この場合、DBMSと応用プログラムとの役割分担の境界をどこにおくか検討する必要がある。

(9) 利用者インタフェース

ゲノム解析に携わるほとんどの研究者は、計算機の利用について習熟しているわけではない。従って、上にあげたものはすべて素人でも利用できるものでなければならない。そのために利用者インタフェースをどのようにするかを考える必要がある。

6. むすび

以上、ゲノム情報の性質およびそれをデータベース化する際の問題点について述べてきた。ゲノム情報のデータベース化には、従来のデータベースの技術では不十分な面が多々あり、データベース研究の立場からも取り組むべき課題が少なくないと思われる。また、本稿では触れることができなかったが、ゲノムの解読には、並列処理や知識処理の技術が必要と考えられている。本稿を契機として情報処理の研究者が一人でもゲノム情報のデータベースやその処理に興味をもっていただければ幸いである。

謝辞

本研究を進めるにあたり、有益なご助言をいただいた東京大学医科学研究所の榊佳之教授に感謝いたします。なお、本研究の一部は平成3年度文部省科学研究費補助金重点領域研究(1)「ゲノム情報」(課題番号 03266101)および平成3年度文部省科学研究費補助金研究成果公開促進費(データベース)の補助を受けている。

参考文献

[BKWMBRKST77] Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, D.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouch, T. and Tasumi, M.: The Protein Data Bank: A Computer-based Archival File for Macromolecular, *J. Mol. Biol.*, 112, pp.535-542 (1977).

[GPKF89] Gray, M.D.P., Patton, W.N., Kemp, J.L.G. and Fothergrill, E.J.: An Object-oriented Database for Protein Structure Analysis, *Protein Engeneering*, Vol.3, No.4, pp.235-243 (1989).

[IsS89] Islam, S.A. and Sternberg, M.J.E.: A Relational Database of Protein Structures Designed for Flexible Enquiries about Conformation, *Protein Engeneering*, Vol.2, No.6, pp.431-442 (1989).

[KMFFSTS84] Kuhara, S., Matsuo, F., Futamura, S., Fujita, A., Shinohara, T., Takagi, T. and Sakaki, Y.: GENAS: A Database System for Nucleic Acid Sequence Analysis, *Nucleic Acids Research*, Vol.12, pp.89-99 (1984).

[KSFTTS91] Kuhara, S., Satou, K., Furuichi, E., Takagi, T., Takehara, H. and Sakaki, Y.: A Deductive Database System PACADE for the Three Dimensional Structure of Protein, *Proc. of the Twenty-Fourth Annual HICSS*, Vol.1, pp.653-659 (1991).

[Lew90] Lewin, B.: *Genes IV*, Oxford University Press and Cell Press (1990).

[MEG89] Mewes, H.W., Elzanowski, A. and George, D.G.: Protein Sequence Databases: Database Management, Data Structures and Data Access, *Biochemical Society Transactions*, Vol.17, No.5, pp.843-845 (1989).

[MTS83] Morffew, A.J., Todd, S.J.P. and Snellgrove, M.J.: The Use of a Relational Data Base for Holding Molecule Data in a Molecular Graphics System, *Computers and Chemistry*, Vol.7, No.1, pp.9-16 (1983).

[Mat91] 松原謙一: ヒトゲノム解析計画の進展と日本におけるプロジェクトについて, 蛋白質 核酸 酵素, Vol.36, No.8, pp.1542-1550, 共立出版 (1991).

[MoT86] Morffew, A.J. and Todd, S.J.P.: The Use of Prolog as a Protein Querying Language, *Computers and Chemistry*, Vol.10, No.1, pp.9-14 (1986).

[NSSUS91] 西川明男, 坂本憲広, 榊佳之, 牛島和夫, 高木利久: 機能部位予測を目的とした遺伝子情報抽出システム, 第29回日本生物物理学会年会講演予稿集 (1991).(発表予定)

[RTNFS86] Rawlings, C.J., Taylor, W.R., Nyakairu, J., Fox, J. and Sternberg, M.J.E.: Using Prolog to Represent and Reason about Protein Structure, in *Third Int. Conf. on Logic Programming* (ed. Shapiro, E.), pp.536-543, Springer-Verlag (1986).

[Raw89] Rawlings, C.J.: Databases, Artificial Intelligence and Knowledge-based Systems for Molecular Biology, *Biochemical Society Transactions*, Vol.17, No.5, pp.851-855 (1989).

[StI89] Sternberg, M.J.E. and Islam, S.A.: A Relational Database of Protein Structure, *Biochemical Society Transactions*, Vol.17, No.5, pp.845-847 (1989).

[WaI90] 渡辺格 監修, 伊藤敏雄 訳: ヒトゲノム解析計画 遺伝情報を解読する巨大プロジェクトの全容, *Newton special issue*, 教育社 (1990).