解　説

学術情報データベースの構成と利用

ゲノムデータベース†

高木　利久††

1. まえがき

1970年代初めの組換えDNA技術に端を発するバイオテクノロジのめざましい発展により，各種生物のゲノム（genome：生物の全遺伝情報を構成する染色体）の構造を決定することが，近年ある程度容易に行えるようになってきた．このような状況の中で，いくつかの生物のゲノム（とくに，ヒトのゲノム）を対象として，その全構造を解析し，その中に含まれるすべての遺伝情報を解読しようとするプロジェクトが，1980年代の終わりごろから米国を中心として各国で始まった[1]．日本では，文部省が1991年度から当面5カ年の計画でヒトゲノムプロジェクトを発足させた．文部省以外にも，科学技術庁，厚生省，通産省，農水省において，これに類するゲノム関係プロジェクトが現在進められつつある[2]．

人間の生物学的な設計図とも呼べるヒトゲノムは，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種類の塩基：アデニン(A)，チミン(T)，グアニン(G)，シトシン(C)によって構成される．ヒトゲノムの場合，先に述べた24本の染色体に含まれる塩基の総数は，約30億対にものぼり，この中には，全体で約10万の遺伝子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[3),4)]．

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．はしごの各段は塩基の対に相当し，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，DNA分子は，どちらか片方の縄，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．

この文字列上には，生物が生命活動を営むうえで必須の分子（たとえば，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．各国におけるゲノムプロジェクトの最終的な目標は，この文字列上の，どの部分に，どのような情報が，どのような形式や規則に従って，格納されているかをすべて明らかにすることである．

後述するように，ゲノムの解読に関わるデータは種類も量も多い．しかも，それらが複雑な関連をもっている．そのため，上記の目標を達成するためには，従来からの分子生物学的アプローチだけでは不十分であり，データベース，知識処理，並列処理といった計算機科学からのアプローチが必要であると考えられている．分子生物学におけるゲノム解析手法によって生み出される多種多様，かつ，大量のデータを，計算機を用いて整理・解析し，その中から生物学的知識を抽出し体系化することが，計算機科学に求められている[5),6)]．

ゲノムプロジェクトは，生物学だけでなく，計算機科学を研究する立場からも興味深い．ゲノムの解読には，従来の計算機科学が対象としてきた問題とは異なる側面がいくつかある．このことは，ゲノム研究が，新たな情報処理技術を生み出す可能性を秘めていることを意味する[7),8)]．

本稿では，ゲノムプロジェクトと計算機科学との接点のうち，ゲノムの解読に関わる情報のデータベース化に焦点を当て，データの種類と性質，データベース化の現状と問題点，研究開発動向について紹介する．

† Genome Databases by Toshihisa TAKAGI (Human Genome Center, Institute of Medical Science, University of Tokyo).

†† 東京大学医科学研究所ヒトゲノム解析センター

2. ゲノム情報

遺伝情報のすべてを解読するには，対象とする生物の文字列（塩基配列）すべてを決定しなければならない．ヒトゲノムの場合，全文字列のうち，遺伝子に相当する部分は数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．そのため，本当に全文字列を決定しなければ，遺伝情報が解明できないかどうかは議論の余地があるが，どのような構成規則からゲノムが作られているかを明らかにするには，ある程度以上の文字列を決定する必要があろう．先に述べたように，ヒトゲノムは約30億の長さの文字列から構成されるが，現在までに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．これでは，あまりに少ない．今後，この文字列（すべて）を実験的に決定する必要がある．

一方，これと並行して，文字列中にどのように遺伝情報が埋め込まれているか，を明らかにしなければならない．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，単に文字列を決定し，その文字情報だけを蓄積するだけでは，その中から遺伝情報を解読することは不可能であろう．いままでの実験や研究で得られている各種のデータや知見を駆使して，遺伝情報を読みとる必要がある．

たとえば，生物的な機能が分かっていない文字列が現れた場合は，それと類似した文字列で機能が既知のものをデータベースから探すことにより（これをホモロジ検索と呼ぶ），その文字列の機能を推定することがよく行われる．また，共通の機能を発現する複数の文字列から特徴的なパターン（モチーフと呼ぶ）を抽出することも遺伝情報を解読する上で必要となる．これらの作業では，異なる生物間（たとえば，ヒトとマウス）での文字列比較も大切である．一方，生物的な機能は，タンパク質の立体構造と深い関わりがある．つまり，遺伝情報を解読するには，塩基配列だけでなく，タンパク質の構造や機能についての情報や各種の生物に関する情報も欠かせないと言える．

実験によって文字列を決定する際にも，レベルの異なる各種のデータを参照することが必要となる．現在の文字列決定技術では，一度に決められる文字列の長さは，数百程度である．そのため，決定した文字列が，どの染色体の，どの位置に存在するかは，染色体の物理地図と呼ばれるものとの対応をとることが必要である．また，その文字列と遺伝子（従来の遺伝学的意味の）とを関係づ

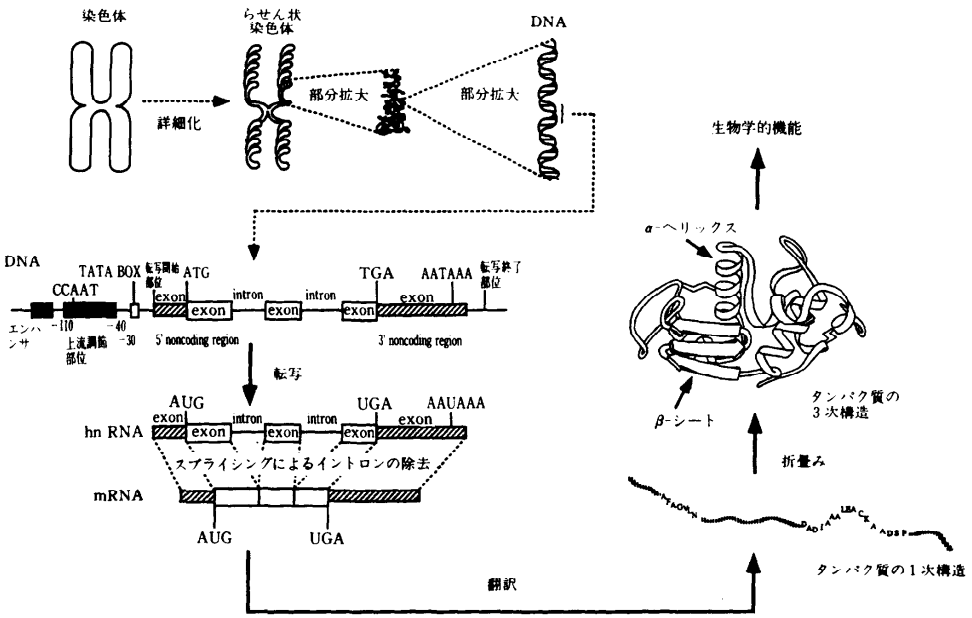

図-1 ゲノム情報の関連図

けるには，遺伝子地図と呼ばれるものとの対応を調べる必要がある．

これらのことは，ゲノムプロジェクトでは各種のデータや生物学的知識を総合的に参照しながら研究を進める必要があることを示している．ゲノムの解読に関わるデータとしては以下のようなものがある．本稿では，これらの情報を総称してゲノム情報と呼ぶことにする．

- 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：遺伝的(リンケージ)解析により作られた遺伝子間の相対的位置に関する地図．
- 物理地図：制限酵素による DNA 断片をもとに作られた塩基配列の物理的位置に関する地図．
- 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（塩基の文字列），2次構造，3次（立体）構造，および文字列の機能に関する情報．
- タンパク質の構造：1次構造（アミノ酸配列），2次構造，超2次構造，3次（立体）構造，およびタンパク質の機能に関する情報．
- 文献情報：各種の構造や機能を掲載した文献に関するデータ．

ゲノム情報の関連の一部を**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おり，ゲノムの解読には，これらのデータを統合して利用できるデータベース環境が必要となる．

なお，ゲノム情報およびその処理全般について，より詳しく知りたい方は，参考文献3)，4)，9)，10)などを参照されたい．

3. データベース化の現状

ゲノム情報のデータベースとしては，研究室レベルで利用するためのラボノートブック，公共的な性格をもつデータベース，ある特定の研究目的のためのデータベース，生物学の知識を体系化した知識ベースなど，いろいろなレベルのデータベースが考えられる．金久らは，これを5つのレベルに分類している[5]．

本章では，この中の公共的なデータベースに的を絞ってデータベース化の現状を紹介する．

まず，ゲノム情報のデータベース化の歴史を概観してみる．

1950年代にタンパク質の構造解析の手法が確立され，アミノ酸配列が決定され始めた．遺伝子に関しては，当時は構造解析の有効な手法はなく解析はあまり進まなかった．1970年代になり遺伝子構造解析手法が開発され，容易に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決定速度は飛躍的に増加した．

塩基配列，アミノ酸配列の急激な増加に対して，データベース化が提案された．アミノ酸配列に関しては，1960年代から，Dayhoff によりジョージタウン大学の NBRF (National Biomedical Research Foundation) においてデータが収集され，1965年に，その成果が出版された．これは1972年からは計算機可読な形となっている．その後，1984年には米国 NIH (National Institute of Health) と NBRF は，PIR (Protein Identification Resource) を設立しデータベースの無料配布を始めた．日本でも1962年よりタンパク質研究奨励会がペプチドおよびタンパク質の情報を収集したデータベースを作成し研究者に提供している．さらに，タンパク質の立体構造についても，1971年に米国のブルックヘブン国立研究所を中心として，PDB (Protein Data Bank) が設立され，タンパク質の結晶構造解析データの収集と配布を行っている．

これに対して，遺伝子では1979年にデータベース化の動きが始まった．欧州では EMBL (European Molecular Biology Laboratory)，米国では NIH が中心となり遺伝子の塩基配列データベースの検討を開始した．1982年には EMBL が EMBL Nucleotide Sequence Data Library Release 1.0 を，米国では NIH が GenBank Release 1.0 を配布し始めた．当初はそれぞれの機関で独自にデータの収集を行っていたが，データの蓄積速度が増えるに従って収集の分担を行うようになった．このような国際的な動きに呼応して，我が国でも1983年に「DNA データバンク運営委員会」が発足し日本における DNA データベースの構築と国際協力の方法について検討が開始された．この方針に従って，1985年に DDBJ (DNA Data Bank of Japan) が国立遺伝学研究所に設けられ，現在では EMBL，GenBank との協力体制のもとに塩基配列データベースの構築を行っている．

表-1 ゲノム情報に関する公共データベース

名称	収集機関	種類	配布媒体	メールサービス	オンラインアクセス
GenBank	ロスアラモス研究所（米国）	核酸1次	磁気テープ CD-ROM	FASTAサーバ ファイルサーバ	オンライン検索 Anonymous FTP
EMBL	ヨーロッパ分子生物学研究所	核酸1次	磁気テープ CD-ROM	FASTAサーバ ファイルサーバ	なし
DDBJ	国立遺伝学研究所（日本）	核酸1次	磁気テープ	なし	オンライン検索
PIR	米国基礎医学研究財団	タンパク質1次	磁気テープ CD-ROM	FASTAサーバ ファイルサーバ	オンライン検索
SWISS-PROT	ジュネーブ大学（スイス）	タンパク質1次	磁気テープ CD-ROM	FASTAサーバ	Anonymous FTP
PDB	ブルックヘブン研究所（米国）	タンパク質3次	磁気テープ	ファイルサーバ	Anonymous FTP
GDB/OMIM	ハワードヒューズ医学研究所（米国）	遺伝子地図	CD-ROM	なし	オンライン検索
MEDLINE	国立医学図書館（米国）	文献情報	CD-ROM	なし	オンライン検索

1980年代に入って，オンコジーンと呼ばれる発癌タンパク質をコードしている遺伝子が発見されたが，タンパク質の機能は不明であった．しかしデータベースとのホモロジ検索により，それまでに知られていたタンパク質と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づくことになった．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た．

表-1は，ゲノム情報に関する公共データベースの中で主要なものについて，データベース収集機関，格納されているデータの種類，検索および配布サービスの形態の現状を示したものである．それぞれのデータベースの詳細については，参考文献11)，12)を参照されたい．ここでは，塩基配列データベースのGenBankを例にとり，その内容を簡単に紹介する．

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．配布の方法は，磁気テープもしくはCD-ROMである．データはテキスト形式である．検索サービスとしては，電子メールによる方法や直接データベースを検索する方法などがある．電子メールによるサービスは，代表的なホモロジ検索プログラムであるFASTAを用いて，利用者から電子メールで送られた配列と類似した配列をデータベースから探し，その結果をやはり電子メールで返すものである．GenBankのデータは，商用の関係データベース管理システムであるSybaseで管理されており，利用者はこのデータベースにも直接アクセスすることが可能となっている．

4. ゲノム情報の性質とデータベース化の問題点

前章でみてきたように，ゲノム情報のデータベースには歴史と実績があり，データベースは分子生物学の研究には，欠かせない道具となっている．しかしながら，現状の公共データベースは基本的にはそれぞれが独立に作られており，これらのデータベースにまたがる検索を行うことは，容易ではない．また，個々のデータベースそれ自体にも問題がある．従来，公共データベースを作成している機関では，受け付けたデータの検査や統合をある程度行っていたが，近年のゲノム情報の爆発的増加に対処しきれずに，それらを未整理のままデータベースに格納する事態になっている．

これらの問題を解決し，ゲノムの研究者がより使いやすい公共データベースを構築するには，生データを検査し，それらを整理・統合するソフトウェアを開発すること，種々のレベルのデータが統合的に扱えるデータベース環境を構築すること，が必要である．また，これと並行して，それぞれの研究目的に応じたデータベースや，研究室レベルのデータを管理するためのデータベース，などの開発を進める必要がある．

以下では，これらの開発を行うのに問題となる

ゲノム情報の性質およびデータベースシステムに要求される技術的課題について，具体的に述べる．

（1） 多様性，階層性，複雑な関連

2. で述べたように，ゲノム情報のデータベースでは，レベルの異なる多種多様なデータを扱う必要がある．たとえば，GenBank 一つをとっても，この中には，DNA の1次元配列データ，その機能に関する記述，それを掲載した文献に関する情報など，性質の異なるデータが含まれている．しかも，1エントリに含まれる配列の長さは，数十から数十万の幅がある．一方，配列データは染色体の地図データ（GDB）をより詳細に表現したものとみなすことができる．このように，現在異なるデータベースとして管理されている GDB と GenBank のデータ間に階層性が存在する．配列データは，タンパク質に関する各種のデータとも関係している．また，ゲノム情報のデータベースでは，このほかに，図形データ（遺伝子地図，物理地図），座標データ（タンパク質の3次構造），階層データ（タンパク質の分類情報），画像データ（各種の実験データ）などを扱うことが必要である．オブジェクト指向データベースやマルチメディアデータベースといった新しいデータベース技術でどこまで対処できるのか，あるいは，より新しいデータモデルが必要なのか，今後検討を加える必要がある．

（2） 重複，誤差，個体差

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，同じデータベース内にも，データの重なりがある可能性があり，重なりぐあ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．たとえば，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，ある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ぎ合わせることが必要となる．データが重複している場合，それらの整合性をどのようにとるかが問題となる．重複したデータを排除することの困難さは，それが個体差に由来するものか，それとも実験誤差によるものかどうか判断できないことによる．DNA の文字列データは，実験の精度からくる誤りを含んでいる可能性がある．また，DNA の解析にどの個体を用いるかによって，実験結果が異なるという問題がある．そのため，研究者によって得られたデータが異なる場合，データベースには，二つの実験結果を併記せざるをえない．逆に言えば，これらのデータ間の違いを調べることにより，生物学的に重要な箇所が分かる場合もある．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，大きな問題である．

（3） 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あげられる．たとえば，文字列の検索一つをとっても，キーワード検索，モチーフ検索，近似ホモロジ検索，厳密ホモロジ検索，などがある．このような検索が必要なのは，塩基配列やアミノ酸配列中に無意味な，あるいは変わっても影響のない部分文字列が含まれていることによる．類似している部分はどこか，あるいは，どの程度似ているか，などが研究上の仮説を立てる際に重要になってくる．これらの多様な処理をそれぞれ高速に行おうとすると，同じデータを，異なるデータ構造で二重三重に格納せざるをえないということにもなる．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，ゲノム情報のデータベースが研究的な色彩を強くもっていることに由来する．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，どの程度まで公共データベースが支援すべきか，どの部分は応用プログラムや専用データベースに任せるか，は大きな課題である．

（4） データ量

表-2 に主要なデータベースに登録されているデータ量を示す．今後これらのデータは爆発的に増加することが予想される．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．しかし，ゲノム研究において，これらのデータどうし

表-2 ゲノム情報のデータ量

名 称	データの種類	データ量
GenBank	塩基配列	エントリ数 65,100 総塩基数 83,894,652
PIR	タンパク質アミノ酸配列	エントリ数 36,150 総塩基数 9,360,161
PDB	タンパク質立体構造	エントリ数 821
GDB	遺伝子地図	遺伝子数 3,123 DNA 断片数 10,243

のホモロジを調べることが，一つの重要な研究手段であり，これは大量の計算パワーを必要とすることから，この問題は軽視できない．高速な検索アルゴリズムの開発やデータベースとスーパコンピュータとの緊密な連携が必要である．

（5） 分 散 化

ゲノム情報は，複雑に関連しており，管理の面から言えば，どこか1カ所で集中管理することが望ましい．しかしながら，公共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，公共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いかに図るかが問題となる．また，計算機ネットワークを介して，分散されたデータベースの更新や利用を行う技術も開発しなければならない．

（6） 処理要求の不明確さ

効率がよく，使いやす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．この際，データベースの柔軟性と効率化のトレードオフが問題となる．また，データ個々についても，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．

5. 研究開発動向

前章までは，おもに公共的なデータベースおよびその統合化に焦点を当て，データベース化の問題点を紹介した．データベースの統合化に関しては，大きく分けて二つの動きがある．一つは，公共データベースを提供する側の動きで，GenBank や PIR といった既存の公共データベースそのものを，根本から再編成しようとするものである．これは，現在米国を中心に検討が進められている．もう一つは，公共データベースを使う側の動きで，既存の公共データベースとは別に，これらを統合的に利用できるデータベースを構築しようとするものである．後者の動向については，のちにもう少し詳しく紹介する．

一方，本稿ではあまり触れなかったが，ゲノムプロジェクトでは，ある特定の研究目的のためのデータベースや研究室レベルで使用するデータベースに対する需要も大きい．これらのデータベースを，ここでは仮に研究用データベースと呼ぶことにする．研究用データベースでは，データの統合化よりもむしろ，高度な検索や推論に重点が置かれる．しかしながら，前章で指摘した問題点の多くは，研究用データベースを開発するうえでも問題になる．このようなデータベースに対しては，現在，演繹データベースやオブジェクト指向データベースからのアプローチがある．

以下では，これらの，ゲノム情報のデータベース化の研究開発動向を簡単に紹介する．

5.1 統合化データベース

ゲノムに関わる各種のデータを統一した環境で利用できることを目的としたシステムは，すでにいくつか開発されており，実用に供されているものも少なくない．代表的なものとしては，IDEAS[13]や GENAS（日本）[14]，IRX（米国）がある．しかしながら，これらのシステムはいずれも検索のインタフェースに関しては統一されているが，前章で述べた意味でのデータの統合が図られているわけではない．

現在，米国 NIH の NCBI (National Center for Biotechnology Information) で開発中の Entrez[15]は，前記のシステムに比べて，統合化が進んだシステムである．Entrez には，GenBank の塩基配列データ，PIR のアミノ酸配列データおよびこれら

に関係する MEDLINE の文献データが格納されている．このシステムの特徴は，データ間に前もって張られた二種類のリンクにある．一つめは，三つのデータベース (GenBank, PIR, MEDLINE) 間にまたがるリンクであり，二つめは個々のデータベース内のデータ間に張られたリンクである．前者のリンクは，配列データとそれを掲載している文献とを対応づけることにより作られる．二つめのリンクは，文献データの場合はキーワードの出現頻度に基づく尺度により，配列データの場合は配列の類似度により，データ間の距離が計算され，それをもとにリンクが形成される．この二種類のリンクにより，たとえば，MEDLINE から文献を一つ見つけると，GenBank 中のそれに対応する塩基配列（一つめのリンク）と MEDLINE 中の関係ある文献（二つめのリンク）が即座に検索できる．このシステムは，パソコン版 (CD-ROM) と X-window 版があり，今秋に公開される予定である．このシステムは，従来のものより統合化が進んでいるが，前章で述べた問題点の本質的な解決になっているものではない．

一方，日本では ICOT の横田，田中らによる研究がある．彼らは演繹オブジェクト指向データベース QUIXOTE を用いて，タンパク質の機能記述を試みている[16]．また，彼らは非正規関係データベース Kappa を用いて，塩基配列およびアミノ酸配列のデータベースを試作している[17]．

5.2 研究用データベース

筆者らは，演繹データベースの手法をゲノム情報の解析に応用し，その有効性と問題点を明らかにすることを目指して，次の二つのデータベースを開発した．

（a） ODS[18]

シグナル配列の研究を支援するために開発したデータベースである．シグナル配列とは，タンパク質によって認識される塩基配列のことであり，遺伝子の発現制御に重要な役割を果たす．ODS では，シグナル配列の長さを考慮して，GenBank から切り出した塩基配列を，長さ 8 の overlapping oligonucleotide に分割して格納している．これにより，データとしての性質が異なる，塩基配列データと生物学的特徴との 1 対 1 の対応づけを図っている．ODS は，推論機能を備えているので，従来のデータベースでは，検索が困難であった高次構造などについても検索を行うことができる．

（b） PACADE[19]

演繹機能を用いてタンパク質の構造（2 次，3 次）を検索するためのデータベースである．PACADE は，構造データを格納した関係データベースと筆者らが開発した推論エンジンからなる．推論機能を用いることにより，タンパク質の疎水性結合や繰り返し構造などの検索が容易に行える．

この二つのシステムの開発により，演繹データベースが，構造記述の面でも，効率の面でもある程度有効であることが確認できた．

筆者らのほかに，演繹データベースや論理プログラミングをゲノム情報のデータベースに適用した例としては，吉田[20]，Morffew[21]，Rawlings[22] らの研究がある．吉田は，ヒトの 21 番染色体の地図を Prolog で記述する試みを行っている．これにより，地図をより柔軟に表現できる．後者の二つの研究は，どちらもタンパク質の構造の記述と検索に，Prolog を利用したものである．

一方，オブジェクト指向データベースを応用した例としては，Gray らの研究[23]がある．タンパク質の構造データ間には階層性がある．たとえば，ヘリックスと呼ばれるタンパク質 2 次構造には，α-ヘリックスや π-ヘリックスなどの種類がある．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．

6. むすび

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対処できない面が多々あり，データベース研究の立場からも取り組むべき課題は少なくないと思われる．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．なお，最近，遺伝情報に関して特許の問題が話題になっている．ゲノム情報のデータベース化という技術的観点からは特許の問題はそれほど重要ではないと思われるので，本稿では取り上げなかった．興味のある読者は，最近の NATURE や SCIENCE などの科学雑誌（たとえば，NATURE Vol. 356, SCIENCE Vol. 254, 255）の記事を参照されたい．

謝辞 本稿を執筆するにあたり，有益なご助言をいただいた九州大学の久原哲助教授，坂本憲広氏，佐藤賢二助手，古川哲也助教授に感謝いたします．

参考文献

1) 米国議会技術評価局編（監修：渡辺，訳：伊藤）：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, p. 173 (1990).
2) 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).
3) Watson, J. D., Hopkins, N. H., Roberts, J. W., Steitz, J. A. and Weiner, A. M. (監訳：松原他)：遺伝子の分子生物学，p. 1163, トッパン (1988).
4) Lewin, B.: Genes IV, p. 857, Oxford University Press and Cell Press (1990).
5) 金久 實，新田克己，小長谷明彦，田中秀俊：知識情報処理技術とヒトゲノム計画，人工知能学会誌，Vol. 6, No. 5, pp. 630-639 (1991).
6) Erickson, D. (訳：中西)：データベース化に悩むヒトゲノム計画，日経サイエンス，1992年6月号，pp. 138-147 (1992).
7) Lander, E. S., Langridge, R. and Saccocio, D. M.: Mapping and Interpreting Biological Information, Comm. ACM, Vol. 34, No. 11, pp. 33-39 (1991).
8) Frenkel, K. A.: The Human Genome Project and Informatics, Comm. ACM, Vol. 34, No. 11, pp. 41-52 (1991).
9) Gribskov, M. and Devereux, J. (eds.): Sequence Analysis Primer, p. 279, Macmillan (1991).
10) 特集：遺伝子情報の解析とタンパク質の構造推定，情報処理，Vol. 31, No. 7, pp. 864-905 (1990).
11) Couteau, J.: Genome Databases, Science, Vol. 254, pp. 201-207 (1991).
12) Nucleic Acids Research, Vol. 19, Supplement, pp. 2221-2249 (1991).
13) Kanehisa, M.: Los Alamos Sequence Analysis Package for Nucleic Acids and Proteins, Nucleic Acids Research, Vol. 10, pp. 183-196 (1982).
14) 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).
15) National Center for Biotechnology Information: Entrez User's Guide (1992).
16) Yokota, K. and Tanaka, H.: GenBank in Nested Relations (Extended Abstract), Proc. Joint Japanese-American Workshop on Future Trends in Logic Programming, pp. 65-74 (1989).
17) Tanaka, H: Protein Function Database as a Deductive and Object-Oriented Database, Proc. Int-Conf. on Database and Expert Systems Applications (DEXA '91), Springer-Verlag (1991).
18) 坂本憲広，高木利久，佐藤賢二，榊 佳之：Development of Overlapping Oligonucleotide Database and its Application to Searching for Signal Sequences over the Human Genome, 情報学シンポジウム，pp. 37-45 (1992).
19) 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. Twenty-Fourth Annual HICSS, Vol. 1, pp. 653-659 (1991).
20) Yoshida, K., Overbeek, R., Zawada, D., Cantor, C. R. and Smith, C. L.: Prototyping a Mapping Database of Chromosome 21, Proc. Genome Mapping & Sequencing Meeting, Cold Spring Harbor Laboratory (1991).
21) 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).
22) Rawlings, C. J., Taylor, W. R., Nyakairu, J., Fox, J. and Sternberg, M. J. E.: Using Prolog to Represent and Reason about Protein Structure, Proc. Third Int. Conf. on Logic Programming (ed. Shapiro, E.), pp. 536-543, Springer-Verlag (1986).
23) Gray, M. D. P., Patton, W. N., Kemp, J. L. G. and Fothergrill, E. J.: An Object-Oriented Database for Protein Structure Analysis, Protein Eng., Vol. 3, No. 4, pp. 235-243 (1989).

(平成4年5月27日受付)

高木 利久（正会員）

1954年生．1976年東京大学工学部計数工学科卒業．九州大学を経て，現在，東京大学医科学研究所ヒトゲノム解析センター助教授．工学博士．演繹データベース，ゲノムデータベースなどの研究に従事．人工知能学会，日本ソフトウェア科学会各会員．