学術研究用たんぱく質データベース PROTEIN-DB†

磯本征雄†† 安岡則武†††
田中信夫††† 松浦良樹††† 角戸正夫†††

学術情報データベースは，データの利用法や親計算機と DBMS との関係など多くの興味ある課題をかかえている．たんぱく質データベースは，これら諸課題の現実的かつ具体的解決手段を明らかにするために，次の2つの目的によってなされた試験研究の成果である．第1の目的は，結晶学分野での将来の幅広いデータ利用に向けたデータベースの雛型システム開発である．第2の目的は，共同利用大型計算機センターの現有システムを活用したデータベース開発のための先導システムの試みである．学術情報一般に共通した多くの特質を持つたんぱく質結晶データを対象とし，汎用 DBMS の活用による共同利用大型計算機センターでの開発と運用の試みは，上記目的達成に好都合であった．本論文では，学術情報データベースとしてのたんぱく質データベースの開発・管理・サービス運用の状況を述べる．

1. まえがき

たんぱく質データベースは，従来より各専門分野で開発研究されてきた学術情報データベースのひとつであり[1]，結晶学分野の雛形システム，他分野の先導システムを目ざして開発された．学術文献情報オンライン検索サービスは既に各所で実施され[2)~5)]，反面で数値データを含む広義の学術情報は，多くの関連諸課題の解決が必要なために，一部でサービス開始またはその準備段階に達したに過ぎない[6),7)]．このような情況下で，これら諸課題の一解決法として，本データベースの開発・管理・運用を紹介することは，有意義であろう．

本データベース開発の特色は，第1に他の学術情報にも共通した特質をもつたんぱく質結晶データを扱ったこと，第2に汎用 DBMS(データベース管理システム）の活用，第3に共同利用施設としての大阪大学大型計算機センターでの開発・管理・運用にある．これらの特色は，たんぱく質データベースの実用化と，今後のための開発経験積み上げの一助を成す目的のためには，幸いに好都合であった．

† Protein Database for Scientific Researchers PROTEIN-DB by YUKUO ISOMOTO (Osaka Unversity Computation Center, Osaka, Japan.), NORITAKE YASUOKA, NOBUO TANAKA, YOSHIKI MATSUURA, and MASAO KAKUDO (Institute for Protein Research of Osaka University, Osaka, Japan.).

†† 大阪大学大型計算機センター

††† 大阪大学蛋白質研究所

* 1978 年度まで東京大学理学部(田隅教授）が日本のサブセンターであったが，1979 年度からたんぱく質研究所(角戸教授）がサブセンターとなった．

さて，たんぱく質データベース開発にあたって，サービス時の情況を配慮し，次の事項を達成目標とした．

① 幅広い適応性：データの利用の多くの局面において，処理効率の良い支援機能を確保する．

② 機能の容易な拡張性：雛型システムから出発し，今後の発展に合せた最新技術の導入や機能の柔軟な拡張・整備の可能なシステムにする．

③ 経費削減と省力化：データベース関与者の多くが，たんぱく質研究者であり，情報サービスのみに人手と経費をかけられない．したがって，開発・管理・運用面での経費削減と省力化が必要である．

④ 明確な標準仕様：データ構造の構成法やデータベース利用法に関して，議論や評価を容易にし，利用者や協力者へのサービス向上に必要である．

本文では，原データの論理構造と利用法，DBMS や親計算機の諸機能との関係などを中心に，たんぱく質データベース開発とサービス環境について述べる．

2. たんぱく質結晶データとその利用法

たんぱく質結晶データは，その内容や頒布・利用面で，他の多くの学術情報にも共通する課題をかかえている．原データは BNL (米国 Brookhaven 国立研究所）で，PDB (Protein Data Bank) の名のもとに収集・頒布事業が進められ[8)]，国内ではたんぱく質研究所が窓口となり*)，BNL と連携して頒布事業を行っている．データの内容は多様であり，利用面でも検索からデータ処理まで多方面にわたる．本章では，このような原データに対するデータベース化までの準備過程について述べる．

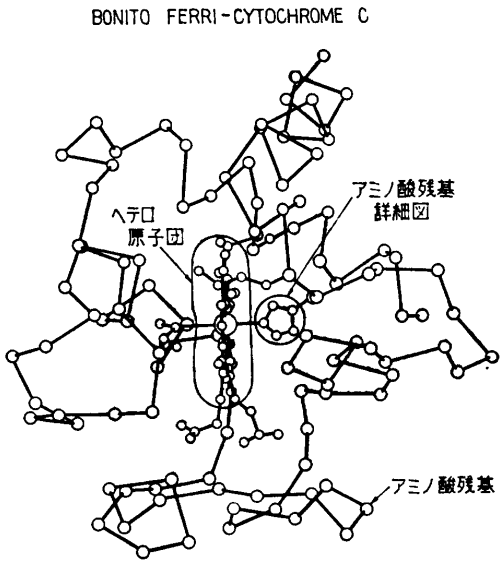


図 1 たんぱく質骨格分子模型
Fig. 1 Molecular backbone model of a protein.

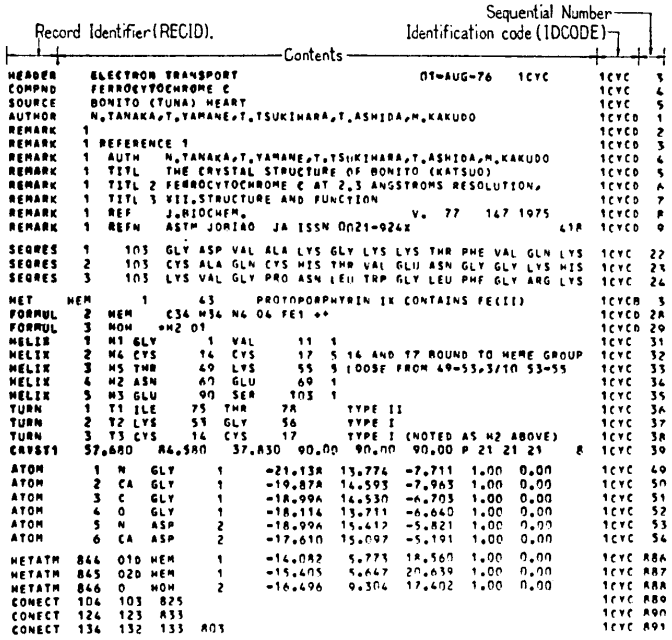


図 2 Protein Data Bank のデータ・フォーマット
Fig. 2 Data format of Protein Data Bank (PDB).

2.1 原データの構成

たんぱく質は，20 種のアミノ酸残基のペプチド結合による鎖とその中心部に位置する亜鉛や鉄などの金属原子を含むヘテロ原子団から成る(**図 1** 参照)．個々の原子座標は，X線回折像の解析結果から得られる電子密度分布に最も良く一致するアミノ酸残基やヘテロ原子団の最適配置の探索で決められる．これらのデータは，1979 年 7 月現在 103 データセット（約 12 メガバイト）であり，1 レコード 80 文字で磁気テープに格納されている(**図 2** 参照)．

たんぱく質結晶データの特徴は，内容の豊富さである．**表 1** は，データの書式や型および内容をもとに，原データ（図 2）のレコードを分類したものである．書誌情報 18%，結晶構造情報 82% の比率である．この分類は，データベース化とそのサービスにおいて，スキーマ記述の構造決定に重要な指針を与え，以下の議論のよりどころとなる．

2.2 網型モデルによるデータ構造の予備的考察

データ構造の構成法として，網型モデルは階層型モデルよりも一般的データ構造の表現が可能であり，また複雑な論理的関係の視覚的表現には関係型モデル[9]よりも有効である．そこで筆者らは，たんぱく質結晶データの全体的構造を考察する目的で，CODASYL 型 DBMS[10]である IDS（網型モデル）を用いて，データベースの予備的構築を試みた[11]．

図 3 は，表 1 をもとに，たんぱく質研究者との協議の結果得られた IDS データベースでのデータ構造の概要を示す．図中四角わく内はレコード名と呼ばれ，原データから再構成されたひとつまたはそれ以上のデータ項目から成る(表 1 参照)．ここでは，データの書式や内容の類似した項目を同一レコードに納めることで，データ構造の単純化に努めた．

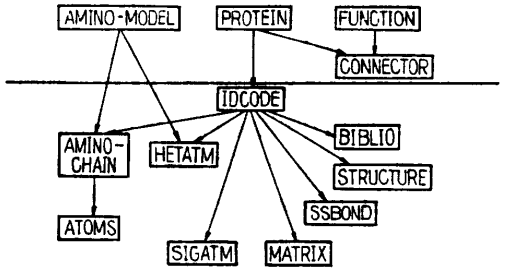


図 3 たんぱく質結晶データ構造の網型モデルによる表現
Fig. 3 Schema of protein crystallo-data by network model.
□ denote record names whose contents are shown in Table 1, and → denote the relations between two records. For example, PROTEIN → IDCODE gives the $1:n$ correspondence of their number of records between PROTEIN and IDCODE.

図中矢印はチェーンと呼ばれ，レコード間の親子関係，あるいはレコード件数における 1 対多の対応関係を示す．たとえば，ひとつのたんぱく質には，構造解

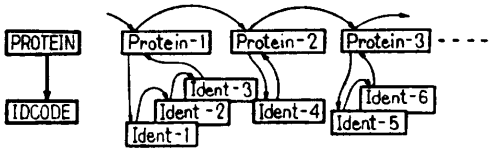
表 1　たんぱく質結晶データの構成とその内容

Table 1　Contents of the protein crystallo-data and their classification.
The contents of record names in Fig. 3 are shown. They are consisted of one or several record identifiers (RECID) in Fig. 2.

分類		Record name in Fig. 3	RECID	Main Contents
書誌情報	概観的内容	PROTEIN	COMPND	たんぱく質名
		FUNCTION	HEADER	化学的機能
		IDCODE	HEADER	当該たんぱく質データのコード
	解説文	BIBLIO	SOURCE	分離された動植物名・微生物名
			AUTHOR	研究者名
			JRNL	関連論文誌
			REMARK	特記事項，参照論文，解析分解能など
			FTNOTE	特定の原子や残基に関連した脚注
結晶構造情報	形状情報	STRUCTURE	HELIX	螺旋形を示す特徴的部位
			SHEET	平板形を示す特徴的部位
			TURN	曲折を示す特徴的部位
			SITE	活性部を構成するアミノ酸残基
			HET	ヘムなど，アミノ酸残基以外の構成成分
	結合情報	AMINO-CHAIN	SEQRES	アミノ酸残基の結合順序
		SSBOND	SSBOND	S-S 結合の部位
			CONECT	ヘムとペプチド鎖の結合部位
	原子座標	ATOMS	ATOM	原子名とそのアミノ酸残基及び原子座標値
		HETATM	HETATM	HET で記載された原子団の原子座標値
		SIGATM	SIGATM	原子座標値誤差の標準偏差
	行列	MATRIX	MTRIX	非結晶学的対称要素を記述する変換行列
			ORIGX	直交座標から著者使用座標への変換行列
			SCALE	直交座標から結晶主軸分率座標への変換行列
			CRYSTL	結晶の格子定数，空間群
	残基	AMINO-MODEL	FORMUL	非標準な部分構造の記述
			etc.	アミノ酸残基やヘムの原子間結合

析技法や測定精度の異なる複数データ・セットがあり，これらは IDCODE で識別されている．この関係は，PROTEIN→IDCODE で表わされる．**図 4** は，これらの関係をさらに詳細かつ具体的に表わしている．

図 3 は，たんぱく質一般にかかわる内容（図中横線より上部）と個々の測定データにかかわる内容（図中横線より下部）に 2 分される．ヘテロ原子団や自然界に実在するわずか 20 種のアミノ酸残基は，複雑な高分子のたんぱく質を構成する基本要素である．個々の測定データの上位に，これらの中の構成原子間結合などの詳細情報を AMINO-MODEL として格納することによって，結晶構造上共通した部分をひとまとめにして取り扱うことができ，以下の測定データのレコード構成の簡単化と冗長性の排除に役立つ．




```
01 PROTEIN.
     02 PROTEIN-NAME          PICTURE X(60).
     98 PROTEIN-IDCODE        CHAIN MASTER.
01 IDCODE.
     02 IDENTIFICATION-CODE   PICTURE A(4).
     98 PROTEIN-IDCODE        CHAIN DETAIL.
```

図 4　IDS によるデータ構造の記述

Fig. 4　Description of data structure by IDS.
The diagram above a broken line shows an extended one about the relation between the records PROTEIN and INCODE. The right hand is the extended diagram for the left hand. A curve with an arrow is a pointer from a record to another. The description below a browken line is a data description by IDS. (01 : record name, 02 : data item, 98 : chain name)

化学的機能はたんぱく質一般に共通した概念である．たんぱく質 (PROTEIN) と化学的機能 (FUNCTION) の間には多対多の対応関係があり，これらは CONNECTOR を仲介にして対応づけられる．たとえば，ヘモグロビンは CONNECTOR を介していくつかの化学的機能に関係づけられる．逆にひとつの化学的機能“親水性”は，やはり CONNECTOR を介して複数のたんぱく質に対応づけられる．この関係は，原データの HEADER の所で各測定データごとに重複して記録されているのにくらべ，一層整理された形式で格納されるので，分りやすい．

たんぱく質分子構造（図 1）を見ると，アミノ酸残基の鎖とヘテロ原子団に 2 分される．これらは，原子間結合の様子が違っている．HETATM には，AMINO-MODEL で約束された順序にヘテロ原子団中の原子座標が格納される．一方，アミノ酸残基の鎖は実際のたんぱく質と同じ順序に AMINO-CHAIN に格納され，各アミノ酸残基ごとの構成原子の座標が ATOMS に格納される．ATOMS 内の原子間結合は AMINO-MODEL にさかのぼれば分かる．

それぞれ IDCOCE の下に並べられたこのほかのレコード名は，上記のものの注釈または補足的内容であり，その内訳は表 1 のとおりである．結局，原データの構造は，以上のような考察を重ねたうえで，網型モデルによって図 3 の中に集約される．

2.3 たんぱく質結晶データの利用法

たんぱく質結晶データの用途を知り，使い易いデータベースを構築するには，たんぱく質研究分野におけるデータの利用法を考察する必要がある．以下，データ利用法の実例を 2〜3 概観する．

たんぱく質研究のひとつは，原子座標の計算機処理による3次元立体構造の解明[12]である．たとえば，アミノ酸残基間のペプチド結合におけるボンド間の捩れ角を統計処理し，法則性を解明する興味深い研究がある．この場合，主鎖上に並ぶ原子の座標をもとにボンドの捩れ角が計算される．別の研究では，グラフィック・ディスプレイ管面上のたんぱく質立体構造図を操作しながらたんぱく質の全体的形状を視覚的にとらえ，既知たんぱく質の結晶構造を見るだけでなく，結晶構造決定の思考実験をくりかえしつつ未知たんぱく質の構造解明に役立てる．たんぱく質構造研究の関心が，大きさや形状にあることから見て，立体構造図[13),14)]の利用は，今後も変わらないであろう．

生化学，医学，薬学部門の研究では，たんぱく質活性部位が興味の対象となり，分子の活性部近傍のデータを頻繁に取り出して利用する[15)]．分子進化学分野では，アミノ酸配列の生物間相互の類似度を調べることで，生物進化の過程での相互の近縁関係を研究する．このためには，アミノ酸残基の一次元配列の特徴に焦点を合せた利用がなされる[16)]．その他，素人向け教育用教材として使われることも少なくない．これら多くの利用法は，サブスキーマとDML（データ操作言語）の利用を通して，具体的データ利用形態として実現される．

3. データの論理構造の記述

論理構造の記述は，DBMS と原データの接点であり，具体的 DBMS に合せて議論される．本データベースでは，転置ファイルの保有，スキーマ記述の見易さ，データロードの容易さ，データベース再構成・再編成の容易さ，プログラミングの容易さ，OS とのインタフェースの良さなどから，INQ[17)]は優利であると考え，最終的にはこれを DBMS として採用した．ここでは説明を簡潔にするために細部は割愛し，図3をもとにして INQ の特徴を有効にいかすために考察した主要な部分のみを述べることとする．

3.1 INQ (DBMS) の概要

INQ の論理構造の記述法は，次の特徴をもつ．

スキーマ記述：データベースは独立なファイルの集合として構成される．論理構造は個々のファイル単位に FDL (File Description Language) によって階層型で記述され，これら FDL の集まりがスキーマとなる．

サブスキーマ記述：サブスキーマは INQ セクションと呼ばれ，1つまたは2つ以上の FDL を結合してつくられ，やはり階層型をなす．FDL の結合の工夫により，INQ セクション段階での多様な論理構造の再構成が可能である．また，異なる INQ セクションを同一プログラムで交互に使うことも可能である．

3.2 スキーマ記述

スキーマ記述は，原データの構造とその利用法，DBMS の特徴を考え合せ，次の考察と手順で決めた．

① FDL への分割：INQ セクション登録時に FDL の再結合が可能なことを有効にいかす．まず，図3のデータ構造を基本的な論理要素に分解する．これにより，多様な INQ セクション再構成の可能性とデータ処理の効率向上を確保する．

② FDL のデータ構造記述：格納データの内容に依存して，検索キー項目を定める．また，論理的基本要素の中での階層構造の記述を具体的に与える．

図5は，手順 ① と ② による FDL の作成例を示す．

図5左側は，図3から抜き出した基本要素であり，右側には INQ データベースの FDL での具体的構造記述を示す．本データベースでは，その他 AMINO-MODEL や，IDCODE とその子レコードを対にして全部で6つの FDL から構成されている．このようにして構成された FDL は，これらの再結合による INQ

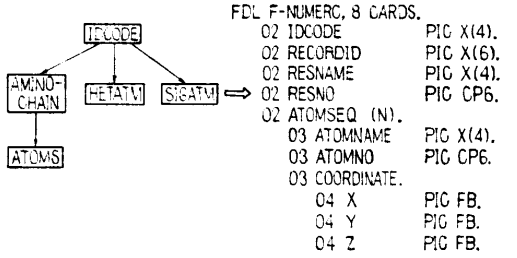


図 5 たんぱく質結晶データの基本要素と FDL のデータ構造記述

Fig. 5 A logical element of the crystallo-data and its data description of FDL for INQ database.
The data for ATOMS, HETATM, and SIGATM are distinguished by the values of RECORDID. The left hand is a logical element which is the part of Fig. 3. The right hand is its data description of FDL.

セクション再構成に際して，IDS のサブスキーマ切り出しよりも多様なデータ構造記述が可能である．なお，データロードは，FDL 単位に独立に，INQ ユーティリティで行われる．

3.3 論理構造の再構成

前節の手順によって管理者側で検討し決定した論理構造が唯一最良であると主張しているわけではない．データベース利用が活発になると，論理構造への要望も多様化するであろう．その場合，本データベースでは 2 通りの論理構造の再構成の方法がある．

軽度の処置としては，既登録 FDL をもとに INQ セクションを組み立てて新たな論理構造を再構成する．この処置はデータの再ロードを伴わないので容易に実行できる．さらに大きな処置が必要な場合には，新規 FDL の登録とデータロードを行う．この場合，既登録 FDL には無関係に実行される．また新規登録のファイル領域は，必ずしも PROTEIN-DB の下に置く必要はなく，各個人ファイルに登録しておいて，利用時に PROTEIN-DB の拡張部分としてアクセスすることも可能である．

4. データ操作言語と応用プログラム開発

バッチ処理利用者には，DML（データ操作言語）が提供される（表 2 参照）．また EUL（エンドユーザ言語）開発など，たんぱく質専門家による支援ソフトウェア開発も，DML に大きく依存する．さて，データ利用法をデータベース・アクセス面から眺めると以下のようになる．

トランザクション処理：格納データを単純再編集して出力装置上に出力する定型処理であり，他データベースとのデータ交換などに必要であろう．条件検索処理：転置ファイルを背景とした複合条件式によるデータ検索である．データ加工処理：実データを読み出し，応用プログラムでこれを加工処理して 2 次情報を生成する．複合処理：条件検索の結果を見ながら必要な場所で必要なデータ移送を行うなどの複雑な処理である．時には，インテリジェント端末を介した処理を行う．多くは未知課題の問題解決手段となる．

表 2 たんぱく質データベース DML 一覧表

Table 2 Data Manipulation Language for the protein database.

機能	DML 命令	機能	DML 命令
条件検索	RETRIEVE SEARCH SORT SAVE AND OR NOT	ファイル管理	OPEN CLOSE CHANGE TSSA TSSR
データ移送	MOVE TABLE KEYLIST	データ更新	STORE MODIFY DELETE
同義語処理	THESAURUS	排他制御	LOCK UNLOCK
FDL 参照	FDL	ジャーナル管理	CLEANPOINT JOURNAL

```
FORTRAN INQ SECTION.
NAME INQFNUM.
        CHARACTER    IDCODE*4
        CHARACTER    RECORDID*6
        CHARACTER    RESNAME*4
        INTEGER      RESNO
*       ATOMSEQ
          CHARACTER  ATOMNAME*4
          INTEGER    ATOMNO
*         COORDINATE
            REAL     X
            REAL     Y
            REAL     Z
C PROCEDURE-1; OPEN THE DATABASE FILE.
      CALL INQ("OPEN"," /F-NUMERC/",1,INQFNUM)
       {name of FDL}———↑
                 {name of INQ SECTION}———↑
C PROCEDURE-2; RETRIEVE ATOMS IN THE AREA.
      CALL INQ ("RETRIEVE",
     &" /IDCODE = '2CHA' AND (( -2.0 < X < 4.0 )
     & AND ( -1.0 < Y < 5.0 ) AND ( 0.0 < Z < 6.0 )) / ")
C PROCEDURE-3; MOVE THE COORDINATES OF THE ATOMS.
  120 CALL INQ ("MOVE","/INQFNUM/ ",MOVE1)
                        └—{name of INQ SECTION}
      IF(INQERROR.EQ."0401") GO TO 200
  140 CALL INQ("MOVE"," /ATOMSEQ/ ",MOVE2)
                        └—{name of group item}
      IF(INQERROR.EQ."0401") GO TO 120
      WRITE(8,1240) RESNAME,ATOMNAME,X,Y,Z
 1240 FORMAT(A6,1X,A4,1X,3(F8.3,1X))
      GO TO 140
  200 WRITE(8,2200)
 2200 FORMAT("END OF DATA")
      STOP
      END
```

図 6 たんぱく質データベース利用のための応用プログラム例

Fig. 6 FORTRAN application program for the protein database. The figure shows INQ section and DML.

本データベースは，これらの処理への適応を配慮しており，いわゆる定型処理システム，データ検索システム，スーパバイザリ・システム[18]，知識システム[19]などで呼ばれているシステムと各々部分的ながらも同じ機能をめざしている．これらは結局，ユーティリティ・プログラムの活用や DML の機能を駆使した，今後のプログラミングの問題に帰着される．

ここでプログラミングの容易さを，FORTRAN プログラムの実例をもって示す．

図 6 のプログラムは，IDCODE が 2CHA のたんぱく質中の，中心（1.0 Å，2.0 Å，3.0 Å）から 3.0 Å 四方にあるアミノ酸残基の全原子座標をファイル（機番 08）上に出力する．図 6 の INQ セクションは，

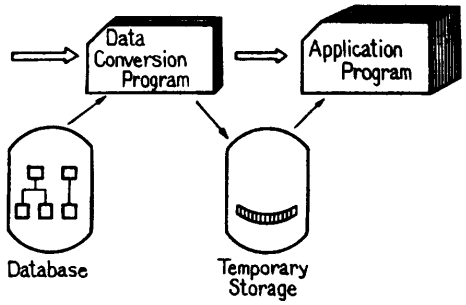


図 7 データベースとプログラム・パッケージの緩い結合
Fig. 7 Loose interface model of application programs to the protein database.

図5の FDL と一対一に対応する．INQ セクション中で左端に＊の付された項目名は，不定繰り返し集団項目と呼ばれ，そのデータ読み出し手順は CODASYL 型のセットとほぼ同様の構造である．ただし，レコード繰り返しの終りは INQERROR＝'0401' によって示される．他の応用プログラムも，大体このような形式で作成されている．

一方で従来より，データベースとは独立に，たんぱく質研究用の多数のプログラム・パッケージが開発されてきた．本データベースとこれらパッケージの間にデータ変換プログラムと中間ファイルを仲介させ，パッケージの変更無しで両者の緩い結合を実現した（**図7** 参照）．この方式は，本データベースで直接サービスできない機能や言語でデータ利用をはかる場合にも有効である．また異種データベースとのデータ交換の機能の一部とも考えられる．データ変換プログラムは，システム構築の際の，諸機能のモジュール化という面から見ても積極的にその価値を評価することができる．

5. エンドユーザ言語

複雑なデータ構造とそれに対する多様な利用形態をもつたんぱく質データベースに必要な EUL 機能は，簡単な単語検索からたんぱく質構造の図示まで多様である．これら多様な機能を実現するには，DBMS の EUL にとどまらず，TSS コマンドや応用プログラムなどのあらゆる手段を駆使して，EUL の関連体系を強化・整理する必要がある．

端末利用者による会話型データベース利用にかかわる本データベースのコマンド関連体系は，次のように大別される（**表3** 参照）．

TSS コマンド群：プログラムの翻訳と実行，ファイルの創生と削除，ファイル間のデータ移送などデータベース・アクセスの準備や後始末に活用される．

表 3 エンドユーザ言語関連コマンド一覧表
Table 3 End User Languages and a part of TSS commands.
TSS user can use directly or indirectly the data of the protein database with the commands of the table.

TSS コマンド（一部分）

コマンド	主な機能	コマンド	主な機能
SCAN	ファイル情報の端末出力	BPRINT	ファイル情報のセンター出力
PERM	一時ファイルの永久ファイルへの転送	RUN	プログラムの実行

基本エンドユーザ言語

コマンド	主な機能	コマンド	主な機能
INQ	EUL の会話開始	FIELD	INQ セクション（サブスキーマ）案内
DONE	EUL の会話終了	?	データベース各種案内
RETRIEVE	論理条件式による条件検索	CHANGE	INQ セクション（サブスキーマ）の切り換え
AND OR NOT	検索されたレコード間の論理演算	SORT	レコードの並びかえ
		CALL	検索テキストの呼び出し
DISPLAY TABLE GRAPH	レコードの内容表示	KEYLIST	キー項目の件数表示
		LET	変数名の定義
SAVE COPY	レコードの保存		

個別エンドユーザ言語

コマンド	主な機能	コマンド	主な機能
SEARCH	単語によるたんぱく質の検索	BIBLIO	書誌情報の端末出力
AMNSEQ	アミノ酸順序によるたんぱく質の検索	PLUTO	たんぱく質構造の図示
CALPHA	C_α 連鎖の座標の一時ファイル出力	FORM	図形の表示形式の変更
RANGE	任意領域内原子座標の一時ファイル出力	ROTATION	図形の回転

基本エンドユーザ言語：DBMS に元来備わっている EUL で，他の INQ データベースとも共通に使われる．これらは，データロード完了後ただちに，利用可能なことが特色である．

個別エンドユーザ言語：たんぱく質データベース固有の EUL であり，本データベースのデータ構造に密接に関連したデータ入出力による能率の良い会話手続きをサービスする．

本データベースの EUL 体系の特徴は，大きく2点に要約される．第1に，管理上の大きな省力化である．TSS コマンドと基本エンドユーザ言語は計算機システム管理者が管理し，個別エンドユーザ言語は本データベース管理者が管理する．この結果，各々の管理区画が小さくなり，管理上の省力化は大きい．しかも今後 INQ を道具とした複数データベースの開設がすすんでも，この体系による個々の管理範囲は明確である．一方，利用者は，これら全体系をひとつのコマンド体系として活用することが可能である．

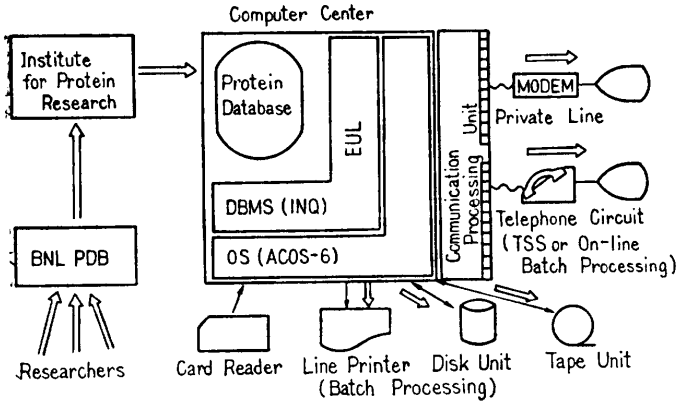


図 8 たんぱく質データベースのシステム構成

Fig. 8 System organization for the protein database. The figure shows architecture of the protein database system and its data procedure. Arrows (⇒) denote the flows of the protein crystallo-data.

第2の特徴は，個別エンドユーザ言語の機能追加が応用プログラム開発の水準で可能なことである．実行形式プログラムをデータベース名 PROTEIN-DB の下に，“書き込み特権をもつ利用者”が登録すれば，TSSコマンドの下で総ての利用者に対してアクセス可能となる．たとえば，図示コマンドを考える．

SYSTEM? PROTEIN-DB/RANGE

は，指定領域内の原子座標を中間ファイル（機番 08）に出力するプログラム RANGE の実行命令である．これに引き続いて図形表示プログラム PLUTO を，

SYSTEM? PROTEIN-DB/PLUTO

で実行すれば，たんぱく質結晶構造詳細図(図1参照)を得る．SYSTEM? は入力促進命令である．

たんぱく質データベースが常にその分野の専門家によって支援され，改善・発展をつづけるためには，このような簡単な手続きによるコマンド登録と呼び出し方式が非常に有効である．また，この登録方式は，計算機システム管理や他のデータベースとは全く独立に管理・運用が可能なので，大型化する計算機システム管理の単純化にも役立つ．

6. たんぱく質データベースのサービス環境

たんぱく質データベースは，EUL による手軽な利用形態から DML による複雑なアクセス技法による高度な計算機利用まで，多様な利用形態に合せた幅広い利用者へのサービスを目的としている．このような多様なデータ処理を可能にするためには，DBMS だけでなく OS の諸機能の支援も必要となる．したがって，本データベースでは，親計算機システム中で，あたかも通常のファイル処理と同程度に OS との十分なインタフェースを確保している(図8参照)．

本データベースのもうひとつの特徴は，利用者が独創的着想の下で独自のデータベース・アクセスとデータ処理技法を工夫できることである．DML による応用プログラム開発だけでなく，新規 FDL とデータロードや INQ セクションの追加・更新も各自の個人ファイル上に容易に登録可能である（第 3.3 節参照）．

データベース管理者にとって，EUL コマンドの提供やスキーマとサブスキーマの広報，データの定期更新と障害対策，ログ情報の管理など多くの任務が必要である．これらは，基本的にはシステム・ユーティリティによる標準手続きに合せて処理される．この点で，支援体系の確立された汎用 DBMS の活用による標準仕様書の存在は，計算機システム管理者とデータベース管理者とデータベース利用者のインタフェースの面での大幅な省力化に役立っている．

現在の情況では，当面必要な諸機能はすべて満たされており，今後は管理者と利用者共にこれら諸特性をいかに有効に活用するかが課題となるであろう．

7. むすび

たんぱく質データベースの利用環境は，計算機システム進歩やデータベースに関与する者の知識や意識の変化とともに変るであろう．したがって，現時点での具体的課題のみにとらわれることのない，柔軟なシステムをつくることが必要であった．本文に示したように，たんぱく質データベースとして特殊な機能や独特の管理・運用体系などは考えず，あくまでも与えられた親計算機システムの枠内で工夫した常識的な方式と道具と場所を用い，広く賛同者の確保しやすいシステムにすることに努め，望めるものなら今後の先導システムの役をも果し得ることを開発目標としてきた．

このような配慮のゆえに，結局以下のような手法による成果を確認することができた．第1に既存の汎用DBMSを活用したことである．これにより，開発・運用の多くの面で経費削減と省力化が可能となった．また，DBMS の汎用化のすすんでいる現在において，類似の後続システムへの貴重な資料提供になる．

第2に多様なサービス機能の確保を重視したことである．本データベースは，情報の統一管理よりはむしろ利用者の多様な要望に応えるものとした．論理構造

の再構成，EUL の強化・改善，既存パッケージと本データベースの緩い結合，データベース管理の計算機システム管理からの分離など，OS との広範なインタフェースも含めた試験開発をすすめてきた．したがって本データベースの適応性は大きいであろう．

たんぱく質データベースは，全国共同利用大阪大学大型計算機センター（ACOS シリーズ 77 NEAC システム 900）および大阪大学蛋白質研究所（ACOS シリーズ 77 NEAC システム S 700）で運用されている[20]．データ量の面では未だ小規模データベースではあるが，機能のうえでは大型データベースに匹敵するものである．現在すでに学内 10 数人の利用者が現われ，全国的広報も開始された．今後データ量の増加と利用者数の増加に伴い，関係者間で様々の要望の具体的実現の努力がなされ，大型の充実したシステムへと発展するであろう．

現在，後続データベース開発[7]の試みがすでに数件すすめられている．本データベースの公共性の高さを反映して，これら後続システムへの多くの参考資料の提供源としても十分に役立つであろう．これまでの経験をふまえ，今後どのように展開してゆくかが，これからの新たな課題である．

最後に，論文体裁に御助言いただきました名古屋工大・川口喜三男教授に感謝致します．

参考文献

1) 文部省特定研究「情報システムの形成過程と学術情報の組織化」第3年次報告，総括班報告 10，3月 (1979).
2) 根岸正光，山本毅雄：オンライン文献情報検索システム・TOOL-IR におけるマン・マシン・インターフェース，情報処理，Vol. 17, No. 5, pp. 402-409 (1976).
3) 中山和彦，及川昭文，上田修一：筑波大学学術情報処理システム——IDEAS/77 を用いたオンラインによる情報検索サービス——，情報処理学会第19 回大会講演論文集，pp. 867-868 (1978).
4) 池田秀人，山本純恭：汎用文献検索システム HUNDRED—思想と特徴—，情報処理学会第 18 回全国大会講演論文集，pp. 559-560 (1977).
5) 渡辺豊英，村尾義和，星野 聰：大学大型計算機センターにおけるデータベース，データベース管理システム研究会資料 12-2，情報処理学会 (1979).
6) 富樫雅文，田中 一：荷電粒子核反応データの収集・蓄積・検索，情報処理学会第 19 回全国大会講演論文集，pp. 879-880 (1978).
7) 弘海原清：単板・関係・網・階層型を含む複合論理データベースについて—学際研究用 DBMS：GEODAS を例として—，データベース管理システム研究会資料 5-1，情報処理学会 (1978).
8) Bernstein T. C. et al.: The Protein Data Bank: A Computer-Based Archival File for Macromolecular Structures, Mol. Biol. Vol. 112, pp. 535-542 (1977).
9) 石井義興：ADABAS モデル，データベース管理システム研究会資料 1-1，情報処理学会 (1977).
10) 西村恕彦ほか：データベースシステム，bit，共立出版 (1977).
11) 磯本征雄：既存の DBMS による蛋白質構造データベースの開発，データベース管理システム研究会資料 4-3，情報処理学会 (1977).
12) 郷 信弘ほか：蛋白質原子座標データとその情報処理，蛋白質核酸酵素，Vol. 23, No. 13, p. 1336 (1978).
13) Motherwell, S.: Cambridge Crystallographic File User Manual (1978).
14) 別府良孝：電子計算機による分子構造の表示，生化学，Vol. 51, No. 1, pp. 24-28 (1979).
15) 成田耕道，崎山文夫：「酵素と高分子触媒」，pp. 3-55，化学同人 (1972).
16) 野田春彦ほか：分子進化学のための情報システム，特定研究「情報システムの形成過程と学術情報の組織化」報告集 1 (1979).
17) 後藤龍男，土井嗣典：INQ (Information Query) について，データベース管理システム研究会資料 3-3，情報処理学会 (1977).
18) 上條史彦：コンピュータ・サイエンス・シリーズ，データベース・システム，産業図書 (1975).
19) 大須賀節雄：知識の表現と利用—知識システムの満たすべき条件—，情報処理，Vol. 19, No. 10, pp. 944-951 (1978).
20) 文部省特定研究「情報システムの形成過程と学術情報の組織化」C-8 班：たんぱく質データベース仕様書，11 月 (1978).

(昭和 54 年 6 月 27 日受付)
(昭和 54 年 8 月 23 日採録)