り転写活性の促進がみられた.

G-d. 遺伝情報分析研究室

日本における DNA データバンクのセンターとして米国，欧州で作成された DNA データベースの導入，国際協力によるデータ入力を進めると同時にデータ解析プログラムの開発，整備，移植を行った.

丸山（遺伝情報研究センター長）は 1987 年 2 月 25 日-28 日ハイデルベルグの EMBL で開かれた "Future Database for Molecular Biology" と題する workshop に参加した.

宮沢は 11 月 12-20 日米国サンフランシスコ郊外で開かれたデータバンクの会合に出席しデータベースの構築に関する議論に参加した.

林田は 4 月に赴任し，データバンク運営に加わるとともに，DNA 配列解析の分子進化学的研究を開始した.

（1） 日本 DNA データバンク (DDBJ) 活動

（i） ニュースレターの発行（丸山・宮沢）: DNA データバンク活動の報告ならびに宣伝のため，1987 年 2 月にニュースレター No. 6 を発行した.

(ii) DNA データベースの導入（宮沢・林田）: 米国から GenBank, NBRF データベース，欧州から EMBL データベースを磁気テープにより取り寄せ，希望者に配布している．配布媒体は GenBank の場合は磁気テープとフロッピーデイスク，その他は磁気テープのみである．配布形態は定期もしくは一時配布である．磁気テープの配布総数は 580 本である．フロッピーディスクの配布枚数は 746 枚である.

(iii) DNA データベースの構築（宮沢・林田）: データ入力をサポートする DDBJ 計算機が 1987 年 3 月に納入され，データのリリースに向けてデータ作成システムの構築を進めてきた．その結果リリース可能となったので，1987 年 7 月 1.0 版をリリースし GenBank, EMBL に配布した．66 エントリー，約 10 万塩基である.

データベースシステムは通常，1) 作成システム（入力，校正，更新）2) 検索システム 3) 解析システムに分けられる．利用者の立場からは，特に 2) 3) が必要とされる．一方一次データ作成を目的とするデータバンクにおいては，組織だったデータ収集，入力，更新を可能にする作成システムが欠かせない．これらのシステムはデータベース管理システム (DBMS) を用い一元的に管理するのが望ましい．しかし作成に時間がかかりデータ入力をはじめている DDBJ にはその余裕がない．我々は作成が比較的容易であるため 3 システムを別々に構築することにした．DBMS を用いる一元的な管理システムの構築は並行して開発していく計画である.

データ作成の手順は，1) DNA データを含む論文の選出，2) Annotators による注釈作業，3) Reviewer による注釈チェック，4) 会社委託によるデータ入力（2 度打ち），5) Annotators による入力エラーのチェック，6) プログラムによるエラーチェック，7) データベースに追加，からなる．林田が Scientific Reviewer としてこのデータ作成

過程を管理している.

現在データ入力に関しては, 米国と欧州のデータバンクである GenBank, EMBL 間で学術雑誌の分担という形で協力がなされている. このような現状から, 丸山 (遺伝情報研究センター長) が参加した 1987 年 2 月 25-28 日ハイデルベルグの EMBL で開かれた "Future Database for Molecular Biology" と題する workshop において, DDBJ に対し GenBank, EMBL との国際協力が要請された. DDBJ としては, 現在各データバンクのデータエントリーが雑誌からの入力である関係上学術雑誌の分担が最も容易であると考え, 日本で出版される学術雑誌を主に担当していくことにした.

(iii-1) DNA データ作成システムの構築: データの履歴管理とプログラムによるエラーチェック (宮沢): データ更新において最も大切なのはデータの履歴管理である. 我々は現在 UNIX オペレーテイングシステムに付属する履歴管理システム, SCCS (Source Code Control System) を用い履歴管理を行っている. SCCS は本来プログラムの履歴管理が目的のため DNA データエントリーにはふさわしくない面もあるが, その利用の容易さから使用することにした. このシステムを使用することにより, 1) バージョン管理 (任意なバージョンの作成, 更新, 回復また更新の理由, 誰が更新したか, その日時の管理) 2) 排他管理 (更新に関する排他性) が可能である. このいずれも, 複数の人によるデータ入力, 更新の管理には必須の条項である.

データ作成過程におけるもう一つの重要な点はエラーチェックである. データを作成するにあたって生物種の学名, 分類等の情報が必要とされる. このため生物分類データベースを作成し維持している. また DNA 配列データからコーディング領域の切り出し, 翻訳等をおこなうプログラムを GenBank より得て移植し, start, stop codons が正しいかどうか, codon frame のチェック等のエラーチェックを行っている. 遺伝暗号表はミトコンドリア DNA で異なる例がよく知られているように, 生物種により若干の違いが例外的に存在するので, 遺伝暗号表のデータベースを作成しそのような遺伝暗号表の違いを考慮している. またこのようなエラーチェックの過程で生成される coding sequence data, protein sequence data は利用価値が高いので各々付随データベースとして配布することにした.

(iii-2) DNA データベース検索システムの構築 (宮沢): 1987 年 3 月より UNIX システムが導入されたため UNIX システムの上で稼働する検索システムを開発してきた. UNIX システムは会話処理向きオペレーションシステムであり, ソフトウェア開発に役立つ多くのツールを所持している. そのような UNIX システムの利点を生かし, 検索に関係する様々な機能をツールとして作成し, それらのツールを組み合わせシェルスクリプトでコマンドを作成する. このような検索システムはポータブルな検索システムとして価値がある. ツールの例は,

- 指定されたエントリーを出力するフィルター
- 指定されたレコードタイプを出力するフィルター
- 指定されたストリング (正規表現) を含むエントリーのエントリー名を出力するフィル

ター
- エントリー名からなるセットに関する and, or, xor フィルター
- Features レコードの記述に従って塩基配列を切り出すプログラム
- 塩基配列からアミノ酸配列に翻訳するプログラム

等である．このようなツールを組み合わせることにより，著者名，論文名，生物種，遺伝子名，キーワード等による検索，特異な塩基配列をもつ遺伝子の検索等が可能である．ファイルシステムとしては保守の簡単なフラットファイルを用いた．今後，処理の高速なプログラムに置き換える作業や，

- 指定された DNA 塩基配列（正規表現）に合致する部分配列をそのエントリー名と共に出力するフィルター

等を作成する予定である．

このような簡易検索システムは，大型，小型，パーソナルコンピューターを問わず UNIX システムなら移植可能であると言う利点を持つ．しかしデータベースの更新，追加，データベースの検索機能がよりすぐれたデータベース，データ作成，検索システムを一元的に管理するようなデータベースを構築するのが望まれる．例えば，Relational data base (RDB) の使用である．現在，GenBank, EMBL, DDBJ 共同で全く同じ RDB を構築しデータの共同入力も含めて共同管理することを計画している．

(iii-3) DNA 塩基配列データ解析プログラムの開発，移植（宮沢，林田）： 解析プログラムの多くは VAX/VMS 計算機上で開発された．これらのプログラムのうちデータベース操作，解析プログラムパッケージ Ideas, UWGCG, PSQ, NAQ, Staden は計算機システムの VAX/VMS (Micro VAX II) にインストールされた．また UNIX システムに移植されたプログラムは系統樹作成プログラム群 Phylip, Staden 等である．（宮沢）

また林田はホモロジーマトリックス表示プログラムを開発した．

(iv) DDBJ 計算機システムのオンライン利用支援システムの構築（宮沢）： 計算機システムのオンライン使用の公開に向けて online information retrieval system を作成し，データバンク活動に関する情報及び計算機システム利用に関する情報のオンラインによる提供を可能にした．またパーソナルコンピューターを端末をして使用するためのプログラムを整備し，希望者に配布した．配布件数は 17 件，フロッピーにして 51 枚である．このような体制ができたので，9 月よりオンライン利用が可能となった．

（2） 核酸の進化速度及び突然変異率の推定（林田）

（i） 突然変異率の性差（宮田*・林田・限*・安永**）： 生殖系列の細胞は成熟までの分裂回数に性差がある．分子進化に関係する突然変異は，主に DNA 複製時のミスによると仮定し，性染色体と常染色体での突然変異率を比較した．理論的には分裂回数の性差が大きい場合，突然変異率は X, Y 染色体において常染色体に比べそれぞれ 2/3, 2 倍と計算される．解析の結果はこの理論値と非常に良く一致し，突然変異率は細胞分裂の回数に

* 九大
** 理研

比例することが示された.

(ii) ウイルスの進化速度の解析 (限*・林田・宮田*): ウイルスの進化は進化機構解明のための格好のモデルであり, 抗原部位の変異速度の情報は医学的にも重要である. 我々は既にインフルエンザA型ウイルス遺伝子の進化速度は核遺伝子の200万倍の速度であることを報告した. 最近, 新しい解析方法を開発し, 数種のウイルスについて進化速度を決定した. その結果, エイズウイルス (HIV-1) の進化速度が最も高く, インフルエンザA型ウイルスの約2倍であることが明らかとなった.

3. 高増殖性癌細胞

ヒト子宮癌細胞 HeLa S3, クローン株	3 株
ラット肝癌細胞	10 株

4. 薬剤抵抗性および修復欠損変異細胞

ヒト・レッシ・ナイハン症由来細胞 (8-アザグアニン抵抗性)	3 株
ヒト・色素性乾皮症由来細胞 (修復欠損)	5 株
シリアン・ハムスター・5-ヨードウリジン抵抗性細胞	5 株
チャイニーズ・ハムスター・8-アザグアニン抵抗性細胞	10 株
チャイニーズ・ハムスター・6-チオグアニン抵抗性細胞	12 株
チャイニーズ・ハムスター・5-ヨードウリジン抵抗性細胞	25 株
チャイニーズ・ハムスター・金属塩抵抗性細胞	10 株

II. 遺伝情報の収集保存

現在 DDBJ に収集されている配布可能な核酸および蛋白質データベースは以下のようである. PGtran は GenBank 35 版からの翻訳データベースである. (Claverie et al., Nature 318, p 19, 1985) SWISSPROT は EMBL フォーマットを用いる蛋白質データベースでそのほとんどは PIR からの変換である. しかし独自に入力も行っている. その他, レトロウイルス (HIV), 免疫関連 (KABAT) の DNA, 蛋白質データベース等利用可能なものがある.

DNA 塩基配列データ:

DDBJ	1 版 (07/87)	66 エントリー	108,970 塩 基
EMBL	13 版 (10/87)	14,397 エントリー	16,023,442 塩 基
GenBank	50.0 版 (05/87)	12,534 エントリー	13,048,473 塩 基
NBRF	31.0 版 (06/87)	2,288 エントリー	4,711,652 塩 基
HIV-N	87.6 版		
KABAT	1973 年版		

蛋白質アミノ酸配列データ:

DDBJ	1 版 (07/87)		
PIR	13.0 版 (06/87)	4,525 蛋白質	1,116,951 残 基
PGtrans	35.0 版 (09/85)	3,107 蛋白質	653,339 残 基
SWISSPROT	5 版 (09/87)	5,205 蛋白質	1,327,683 残 基
HIV-N	87.6 版		
KARAT	1987 年版		

以下は各データベースの簡単な収集内容である. なお, 図はこれまでの各データベースの収集数の変遷を描いたものである. EMBL の収集塩基数が 1985 年中頃急増しているが, これは GenBank と相互にデータ交換を始めたためと思われる.

1. GenBank Release 50.0

グループ	エントリー数	塩基数
霊長類	1,565	1,930,507
ゲッ歯類	1,792	1,713,598
哺乳類	372	394,441
脊椎動物	545	518,518
無脊椎動物	775	676,378
植物	782	1,013,120
オルガネラ	449	881,664
バクテリア	1,027	1,482,299
RNA	687	76,128
ヴィールス	1,211	1,831,351
ファージ	179	969,959
合成	253	79,232
無注釈	2,897	2,154,278

2. EMBL Release 13

グループ	エントリー数	塩基数
人工	223	91,717
クロロプラスト	203	520,683
遺伝因子	58	51,617
ミトコンドリア遺伝子群	367	434,795
原核生物	1,501	1,846,153
ヴィールス／ファージ	1,567	2,657,169
真核生物	7,467	8,133,595
その他	25	48,744
無注釈	2,986	2,239,005

3. NBRF Release 31.0

グループ	エントリー数	塩基数
真核生物	2,288	2,449,858
哺乳動物	779	1,396,208
植物と真菌類	259	637,221
真核生物ヴィールス	400	1,352,139
原核生物	460	694,144
バクテリオファージ	64	215,511
動物ヴィールス	351	1,221,440
植物ヴィールス	49	130,699
大腸菌	225	382,423

真菌類	176	259,209
人類	285	658,168
ミトコンドリア	61	200,115
クロロプラスト	30	31,995

4. PIR Relsase 13.0

グループ	エントリー数	残基数
真核生物	2,841	579,525
哺乳動物	1,575	342,985
植物	327	62,610
真菌類	157	48,262
原核生物	741	167,591
動物ヴィールス	625	284,818
植物ヴィールス	67	28,961
バクテリオファージ	252	56,754