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．

（1） DNA データベースの導入（宮沢）：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ィスク，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．今年度の配布実績は以下のようである．

データベース		配布数 (12 月 18 日現在)		
DNA データベース:			VAX/VMS 版	フロッピー
GenBank	40 版	1		2
	42 版	12		
	44 版	20		13
EMBL	8 版	16		
	9 版	15		
NBRF	27 版	9		
	28 版	13	1	
	29 版	13	8	
蛋白質データベース:				
NBRF (PIR)	8 版	19	1	
	9 版		1	
	10 版	20	8	
PGtran	35 版	16		

磁気テープの配布総数は 209 本である．フロッピーディスクの配布枚数は 234 枚である．

（2） DNA データベースの構築（丸山・宮沢）：データ入力のための予算（データ入力委託費）が今年度より付き，データ入力を開始した．フォーマットは GenBank に準じるものを用い，データ注釈者のためのマニュアルを宮沢が作成した．データ注釈は研究所内の大学院生に依頼し，scientific reviewer として各大学院生の指導教官の協力を仰いだ．またデータ注釈のフォーマットが統一されるよう最終のチェックを宮沢が行った．

データ入力の手順は以下のようである．

（i） 学術雑誌の選択：データ入力に関する国際分業に備え，日本で出版される学術雑誌を優先的に取り上げた．それ以外の雑誌は，関連論文が多く見い出される雑誌を選んだ．

（ii） 関連論文の選出：定期的に出版される学術雑誌から関係論文を選出することはそれ自体手間のかかる仕事である．今年度はデータ注釈者は全て研究所内の大学院生のため，

分子遺伝研究部門石浜教授の協力を得，雑誌毎に担当者を決め，論文を選出してもらった．将来，外部にも注釈者を依頼する場合には，文献情報誌（蛋白質研究奨励会より出版されている Peptide Information）の利用も考えている．

(iii) データ注釈/データコーディング：文献を読み DNA 配列に関する有用な情報（プロモーター，オペレーター部位，リボゾーム結合部位等）を得，一定のフォーマットでコーディングする．このようなデータ注釈作業はデータの質を決める重要な作業であり，データ入力で最も時間のかかる部分である．特にデータ注釈の統一に努力した．

(iv) データ入力：コーディング用紙からのデータ入力作業は，誤りを少なくするため 2 度入力し，更に塩基配列部分は音声出力を用いてチェックした．以上のようなデータ入力作業は日立ソフトウェアエンジニアリングに委託された．

(v) 入力データのチェック：入力データの最終チェックは各データ注釈者に依頼した．

データ入力の国際協力に関しては，要望が強く，そのためのミーティングが 1987 年 2 月ドイツで開催される予定である．GenBank, EMBL の関係者と共に丸山が参加する予定である．このミーティングではデータ注釈のフォーマットの改良に関する話し合いも予定されている．またデータ入力に関する協力を要請すべく，1986 年秋，丸山は渡米の途中 GenBank 関係者と会談した．

(3) DNA データベース検索システムのデザイン（宮沢）：DNA データベースの使用にあたって検索システムが必須であることは言うまでもない．1987 年 3 月より新計算機システムとして UNIX システムが導入される予定のため，UNIX システムの上で稼働する検索システムが必要とされる．UNIX システムはこれまで研究所にあったタイプのオペレーションシステム (MSP) とは違い，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とを計画，開発している．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と思われる．

ツールの例は，

—指定されたエントリーをエントリー名と共に出力．
—指定されたレコードタイプをエントリー名と共に出力．
—指定されたストリング（正規表現）を持つレコードをエントリー名と共に出力．
—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．
—フォーマット変換

等である．このようなツールを UNIX システムにあるツール (sort, uniq...) と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

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

L. ウ ズ ラ (*Coturnix coturnix japonica*)

1. 突然変異系統

アルビノ

2. 閉鎖群

野生起原群，家禽化群

II. 遺伝情報の収集保存

DNA データバンク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．

DNA 塩基配列データ：

GenBank	44.0 版	8,823 エントリー	8,442,357 塩 基
EMBL	9 版	7,630 エントリー	7,813,214 塩 基
NBRF	29.0 版	1,988 エントリー	3,686,016 塩 基
DDBJ (プレエントリー)		約 4,400 エントリー	約 3,700,000 塩 基

蛋白質アミノ酸配列データ：

PIR	10.0 版	3,800 蛋 白 質	890,703 残 基
PGtrans	35.0 版	3,107 蛋 白 質	653,339 残 基

(1) GenBank

グループ	エントリー数	塩 基 数
1. 霊 長 類	1,028	1,240,779
2. ゲッ歯類	1,272	1,111,622
3. 哺 乳 類	245	244,554
4. 脊椎動物	474	400,509
5. 無脊椎動物	605	435,280
6. 植 物	594	643,365
7. オルガネラ	368	485,666
8. バクテリア	749	1,031,546
9. R N A	637	69,232
10. ヴィールス	1,093	1,517,025
11. ファージ	160	271,817
12. 合 成	224	72,029
13. 無 注 釈	1,374	918,933

30,000 塩基以上よりなる遺伝子群

エントリー名	グループ名	塩基数
HUMTPA	霊長類	36,594
HUMFIXG	霊長類	38,059
HUMHBB	霊長類	73,360
AD2CG	ヴィールス	3,5937
T7	ファージ	39,936
LAMBDA	ファージ	48,502
EBV	EBV	172,282

(2) EMBL

グループ・	エントリー数	塩基数
人工	182	68,540
クロロプラスト	149	153,786
遺伝因子	54	43,857
ミトコンドリア遺伝子群	307	346,721
原核生物	1,065	1,130,637
ヴィールス／ファージ	1,195	1,689,681
真核生物	4,668	4,364,797
その他	10	15,195

(3) NBRF

グループ	エントリー数	塩基数
真核生物	1,183	1,793,002
哺乳動物	637	1,091,396
植物と真菌類	246	336,726
真核生物ヴィールス	316	1,069,842
原核生物	435	647,351
バクテリオファージ	54	175,821
動物ヴィールス	274	965,986
植物ヴィールス	42	103,856
大腸菌	212	358,297
真菌類	160	223,976
人類	229	517,048
ミトコンドリア	58	187,015
クロロプラスト	34	56,168
計	1,988	3,686,016

(4) PIR

グループ	エントリー数	残基数
真核生物	2,380	449,474
哺乳動物	1,329	273,569
植物	217	31,839
真菌類	139	40,682
原核生物	664	148,184
動物ヴィールス	483	223,779
植物ヴィールス	53	22,233
ファージ	221	47,731
計	3,800	890,703