斎藤成也（1993）
「モンゴロイド」No.16. 29-35頁

遺伝子頻度データベース『FREQ』の開発と人類集団の遺伝的近縁関係の分析

斎藤成也

筆者は遺伝子頻度データベース『FREQ』を、1990年度から開発している（斎藤成也「遺伝子頻度データベースの開発」本誌No.11）。また、人類集団の遺伝的近縁関係の分析を初年度の1989年度から行なっている。本稿ではこれら二つのことについて述べる。

遺伝子頻度データベース『FREQ』は、遺伝子頻度（gene frequency）の"frequency"から取った名前であるが、今回その試作版を作成した。データ量が全体としておよそ1メガバイトになったので、2HDのフロッピーディスク1枚に納めることができた。この試作版の概要を以下に述べる。データベース『FREQ』には3種類のファイルがある。そのうち、『集団ファイル』と『遺伝子座ファイル』は、ちょうど1年前に行われたシンポジウムでもすでに説明したとおりだが、今回は『文献ファイル』というのも新たにつけ加えた。これ自体は前から集団ファイルを作成するための補助資料として入力を続けてきたもので、遺伝子頻度データを発表している論文のリストである。リストには2種類あり、一つはRoychoudhury & Nei（1988）のデータブックで引用されている全文献のリスト、もうひとつは私が独自に収集した論文のリストである。後者は、現在600強のリストをすでに入力してある。

FREQ試作版フロッピーディスクに入っているファイルはREADME．DOCというファイルに一覧がある（表1）。PLOT1というのは遺伝子頻度の地理的分布を自動的に作図するプログラムである。その使用法、それからそのプログラムが参考にするデータファイルが入っている。POP、REF、LOCUSという三つのディレクトリは、それぞれ集団ファイル、文献ファイル、遺伝子座ファイルに対応する。

遺伝子座ファイル（LOCUS）は、赤血球酵素はRBCというディレクトリ、血清タンパクはSERUMというディレクトリ、血液型はBLOODというディレクトリに分かれている。いまのところ、この三つに分けているが、近い将来は関係諸氏の協力を得て、たとえば1991年に行なわれた第11回HLA国際ワークショップのデータを加えることを考えている。

赤血球酵素に関しては、いまのところ8個の遺伝子座のデータだけを入れてある（表2）。1991年も同様のものを出したが、集団の数ははるかに多くなっている。

表1．FREQ試作版フロッピーディスクに入っているファイルのリスト

ルートディレクトリ¥		
COMMAND	COM	MSDOSのシステムファイル
README	DOC	FREQの概要を説明したファイル
PLOT1	DOC	プログラムPLOT1の使用法
PLOT1	EXE	遺伝子頻度の地理的分布を自動的に作図するプログラム
DATA	<DIR>	プログラムPLOT1が参照する世界地図データファイル
I_DATA	<DIR>	プログラムPLOT1が参照する遺伝子頻度データファイル
POP	<DIR>	集団ファイル
REF	<DIR>	文献ファイル
LOCUS	<DIR>	遺伝子座ファイル

血清タンパクについては、表3に示したように20個近い遺伝子座のデータを入力している。

血液型に関しては最近始めたばかりで、とりあえず5つの遺伝子座(ABO、AU、CO、CS、DI)のデータを入れてある。

LOCUSファイルが具体的にどういうものかというと、図1はPGDという赤血球酵素遺伝子座のデータの実際の中身の一部である。まず第1行に、『Table 64 Phosphogluconate dehydrogenase : PGD』とあるが、このテーブル番号はRoychoudhury & Nei (1988) のそれに対応している。それからA、C、Xという三つの対立遺伝子名があり、次に、各行の最初の二つの数字は、経度と緯度を与えている。これは、世界地図を引っ張りだしていちいち各集団を世界地図上でどの経度・緯度を与えるかというのを調べていくわけである。国名で与えられている場合はとりあえず首都の位置を与えているが、集団名などの場合はそうもゆかず、ここの入力作業はかなり難航している。3番目のカラムは世界的な地域を示しており、Eurはヨーロッパ、Afrはアフリカなどとなる。4番目以降は、集団名、遺伝子頻度、標本数である。RNというのはRoychoudhury & Nei (1988)からの引用であることを意味する。将来はそれ以外のものも付け加えていく予定である。遺伝子頻度はこの全部の頻度の合計が1になる性質があるので、プログラムをかけて合計が1になることをチェックしている。

次に付属のプログラム、PLOT1について説明する。このプログラムは、A05班の新見康永・小林豊両氏のグループにお願いして作っていただいた。今回説明するのはフロッピーディスクのFREQ試作版に入れたプログラムの改定版である。

このプログラムは、世界地図のデータファイル及び遺伝子頻度のデータファイルが必要である。

最初のプログラムを走らせると、「地図の図法はどれにしますか」と聞いてくる(図2)。このように同じ一つの地図情報ファイルをもとに、数式を使って変換して地図をミラー図法、サンソン図法、正射図法という3種類の図法で表示することができる。

1番のミラー図法を選ぶと「使用ファイルを選んでください」と聞いてくる(図3)。これらは、RBC(赤血球酵素)、SERUM(血清タンパク)、血液型(BLOOD)という3つのディレクトリに対応しているわけである。

たとえば1番の赤血球酵素を選ぶと、「表示するデータを選んでください」と聞いてくる(図4)。1番～8番の遺伝子座位を選んでくださいと聞いてくるわけである。ここで仮に1番(Acid Phosphatase)を選ぶと、まず世界地図をかきだす。ユーラシア、アフリカから始まり、オセアニアが出る。次に北アメリカを描いて、最後に南アメリカが出る。パソコンの能力にもよるが、ここまでに1分間から2分間ぐらいかかるかもしれない。終わると次に遺伝子頻度分布の円グラフを作成するが、集団の数が多いと計算に時間がかかる。この酸性フォスファターゼの場合は5分間くらい待たなければならない。

数分間たつと、このように突然多数の円グラフが出現するわけである(図5)。円グラフの数が少ない場合、たとえば2～3個だと世界地図を描き終わっ

表2．赤血球酵素遺伝子座の遺伝子頻度データファイルのリスト

AK1	TAB
ACP1	TAB
GPT1	TAB
ESD	TAB
PGD	TAB
PGM1	TAB
GL01	TAB
ADA	TAB

表3．血清タンパク遺伝子座の遺伝子頻度データファイルのリスト

C1R	TAB
F13A	TAB
F2	TAB
BG	TAB
LD	TAB
ATP	TAB
AP0A4	TAB
CP	TAB
F13B	TAB
LP	TAB
P1	TAB
PISUB	TAB
AIISG	TAB
AP0E	TAB
AT3	TAB
AG	TAB
C2	TAB
C3	TAB

Table 64. Phosphogluconate dehydrogenase: PGD				(A, C, x)				
4.12	50.50	Eur	Belgium	0.984	0.017	0.000	500	RN
23.18	42.40	Eur	Bulgaria	0.988	0.011	0.000	138	RN
14.25	50.50	Eur	Czechoslovakia	0.988	0.012	0.000	330	RN
12.34	55.43	Eur	Denmark	0.980	0.020	0.000	1574	RN
25.00	60.08	Eur	Finland	0.968	0.032	0.000	282	RN
2.20	48.52	Eur	France	0.995	0.005	0.000	608	RN
13.25	52.32	Eur	Germany	0.980	0.020	0.000	1162	RN
19.03	47.30	Eur	Hungary	0.987	0.013	0.000	116	RN
-5.50	54.40	Eur	Ireland	0.986	0.014	0.000	1797	RN
-21.64	64.09	Eur	Iceland	0.979	0.021	0.000	1069	RN
12.30	41.53	Eur	Italy	0.946	0.054	0.000	193	RN
4.54	52.52	Eur	Netherland	0.979	0.021	0.000	601	RN
14.29	52.54	Eur	Poland	0.974	0.026	0.000	213	RN
-3.43	40.25	Eur	Spain	0.984	0.016	0.000	188	RN
18.95	59.20	Eur	Sweden	0.981	0.019	0.000	412	RN
0.10	51.30	Eur	England	0.979	0.021	0.000	4567	RN
93.55	54.22	Eur	USSR	0.962	0.038	0.000	1553	RN
13.12	32.58	Afr	Libya	0.952	0.048	0.000	146	RN
5.55	23.50	Afr	Algeria	0.987	0.013	0.000	224	RN
31.15	30.03	Afr	Egypt	0.962	0.038	0.000	239	RN
32.30	15.39	Afr	Sudan	0.976	0.024	0.000	287	RN
16.02	8.36	Afr	Chad	0.995	0.005	0.000	207	RN
-7.59	12.40	Afr	Mali	0.836	0.164	0.000	283	RN
-17.24	14.38	Afr	Senegal	0.952	0.048	0.000	756	RN
-10.46	6.20	Afr	Liberia	0.941	0.059	0.000	485	RN
-4.01	5.19	Afr	Ivory Coast	1.000	0.000	0.000	186	RN
3.28	6.27	Afr	Nigeria	0.945	0.055	0.000	64	RN
11.31	3.51	Afr	Cameroon	0.954	0.042	0.004	284	RN
38.42	9.03	Afr	Ethiopia	0.865	0.135	0.000	100	RN
32.35	0.19	Afr	Uganda	0.964	0.036	0.000	155	RN
32.50	-5.02	Afr	Tanzania	0.941	0.054	0.005	211	RN
13.12	-8.48	Afr	Angola	0.917	0.083	0.000	109	RN
18.28	-33.56	Afr	Southern Africa	0.966	0.034	0.000	119	RN
-150.00	61.10	NAm	Alaska	1.000	0.000	0.000	170	RN
-75.45	45.25	NAm	Canada	0.990	0.010	0.000	136	RN
-122.20	47.35	NAm	USA	0.982	0.018	0.000	547	RN
-99.10	19.25	CAm	Mexico	1.000	0.000	0.000	85	RN
-90.22	14.38	CAm	Guatemala	0.983	0.017	0.000	205	RN
-84.04	9.55	CAm	Costa Rica	0.897	0.103	0.000	286	RN
-79.30	8.57	CAm	Panama	0.946	0.054	0.000	402	RN
-62.25	23.07	CAm	W Indies	0.539	0.058	0.003	308	RN
-68.10	-15.30	SAm	Bolivia	0.986	0.001	0.003	316	RN
-68.45	-10.59	SAm	Brazil	1.000	0.000	0.000	363	RN
-75.26	-50.46	SAm	Chile	0.959	0.041	0.000	176	RN

図1．遺伝子座ファイルの例（赤血球酵素PGD）

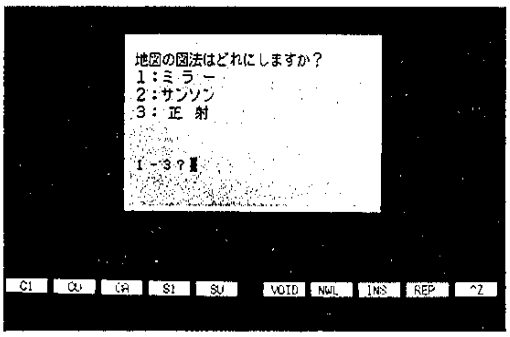

図2．プログラムPLOT1の画面（その1）

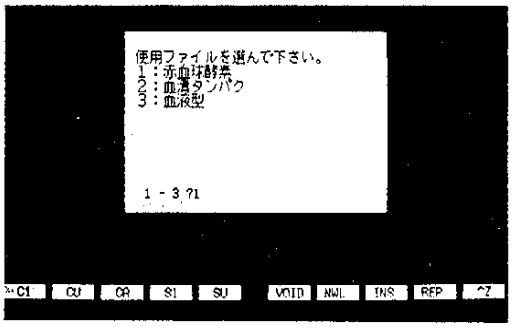

図3．プログラムPLOT1の画面（その2）

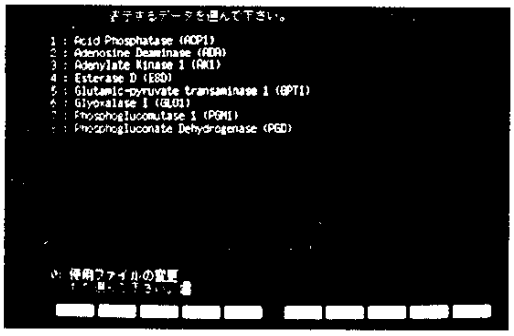

図4．プログラムPLOT1の画面（その3）

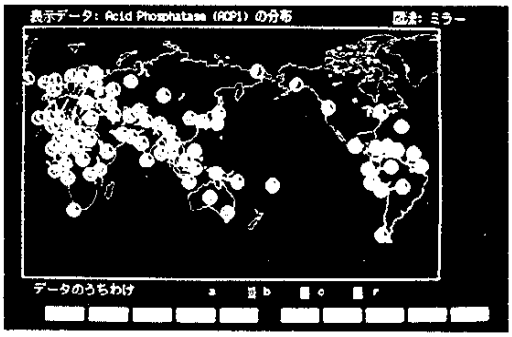

図5．プログラムPLOT1の画面（その4）

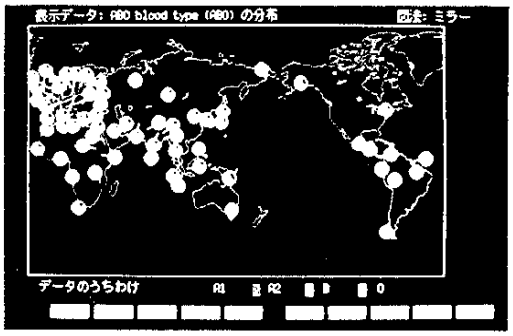

図6．プログラムPLOT1の画面（その5）

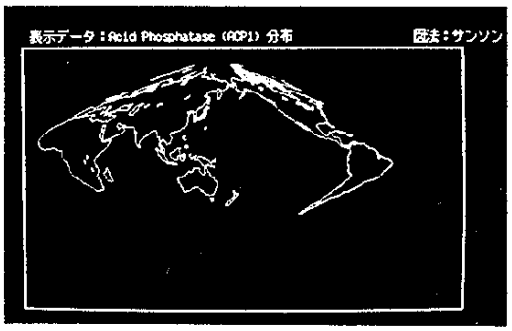

図7．プログラムPLOT1の画面（その6）

た後、瞬間的に出てくるが、これだけ多数の円グラフを表示しようとすると、かなり計算がめんどうになるらしい。対立遺伝子の名前はa、b、c、及びr(rare variant、まれな変異型遺伝子)である。

最終行にRETRY(繰り返しますか)という質問があり、イエス・ノーを聞いてくる。ここでYと押すと、また「使用ファイルを選んでください」(図3)の画面へ戻り、1番、2番、3番と選べる。

血液型を選ぶと、5種類の遺伝子座の一覧表が出てくる。たとえばABO血液型を選ぶと、再び世界地図を描き、その後に円グラフが表示される(図6)。

ABO血液型は昔から知られているように南米の大部分ではOの遺伝子頻度が100%であることがわかっている。このように集団が増えてくると、次々に重なってきてわからなくなってくるので、データが増えた場合にいかに世界地図を効率よく表示するか、ということもいま検討していただいている最中である。

いままでのは全部ミラー図法であるが、サンソン図法を選ぶと、少しひしゃげたような投影を行なう(図7)。そのほか正射図法もあるが、基本的にはミラー図法だけでいいのではないかと思っている。将来はこのミラー図法の世界地図の画面データを一つのマップのファイルにしておくと、それぞれ描かずに瞬間的に画面に表示できるようになると期待している。

次に集団ファイルの説明に移る。集団ファイルについては、1991年の発表ではテキストファイルということであった。初年度だったのでいろいろ試行錯誤していたが、基本的には数値ファイルで、それをコンピュータの入力ファイルにできるような形にしておいたほうがいいと考え、いま試験的に表4のような幾つかの集団を入れている。たとえば、1番上がスリランカのヴェッダ、シンハリ、バリ島のバリ人、フィリピンのアエタ(ネグリト)である。

表4．集団ファイル一覧

VEDDA___E	POP
SINHAL__E	POP
BALI_____	POP
AETA_____	POP
FILIPINO	POP
MONGOL__C	POP
KOREAN__C	POP
ZHUANG__C	POP
MANUS___	POP
SAMOANS__	POP
IFUGAO___	POP
LEPCHA___	POP
TAGALOG__	POP
VISAYAN__	POP
AINU____O	POP

集団ファイルの内容をMANUS_____.POPファイルを例にとって説明する(図8)。ニューギニアの北に浮かぶマヌス(Manus)島という島があるが、そこに住むメラネシア人集団のデータであり、1972年にHH(Human Heredity)にMalcolmらが発表したものである。もう20年前のデータであるが、その時点でABO血液型から始まって、赤血球酵素のAK、G6PDといったものが調べられている。これは2行で一つのセットになっており、たとえばABOでは101という番号が最初のカラムにあるが、これは私がつけた座位の認識番号である。138というのは標本数、その後に遺伝子座の名前、およびその対立遺伝子の名前が続く。その次の行が実際の遺伝子頻度のデータである。このように2行で一つのブロックになっている。このようなデータファイルを集めて、いろいろな集団の遺伝子頻度データを複合させたものを出力する。その出力ファイルが、逆に入力ファイルとなって遺伝距離が計算される。遺伝距離を推定した後、その距離行列から集団間の近縁関係を推定する、というプログラムシステムを大体作りあげた。

最後に文献ファイルであるが、これはRoychoudhury & Nei (1988)の文献リストファイルを全部入れたものをまずARUNというディレクトリに入れた。Roychoudhury氏のファーストネームはArunなので、その名前にした。AからBに始まって、最後にZまでいく。かなり大きいファイルなので、小さく区切っている。これらをたとえばエディターで合体してもかまわないし、全部プリントアウトしてもいいかもしれない。

データベースにするというのは機械可読型になるので、自由に検索ができるわけである。単に本をめくって一生懸命探すのではなく、たとえばある遺伝子のデータを誰が発表したか忘れてしまったという場合、あるいはあの人の論文だったがどれだったかというときに、エディターやワープロソフトを使えば比較的簡単に検索できる。そんな意味でこれら文献情報を入力したわけである。

次に、私自身が収集している遺伝子頻度データに関しては、DBASEIIIというパソコン用のデータベ

ースソフトを使い、ID、雑誌名、巻、ページ、第1著者、出版年、号、第2著者、論文タイトル、集団、遺伝子座、というフィールドを決めて入力を進めている。先ほど述べたように、これが600件以上になっている。DBASEIIIが無い場合のため、これで作成したファイルをMSDOSで普通に読めるようASCIIファイルに変換したものをつけてある。ID番号のほかに、雑誌名、巻、ページ、第1著者、出版年、タイトルというようになっている。以上が遺伝子頻度データベース『FREQ』の試作版の紹介である。

次に、人類集団の遺伝的近縁図の作成の話に移る。主にオセアニア、南北アメリカ、アジアといったモンゴロイド集団を中心に世界中の50人類集団を調べた（図9）。

たとえばMN血液型のM遺伝子頻度をみると、遺伝子頻度が0から1まで分布するが、オセアニアの集団はMの頻度が低い、あるいは中南米とか北米といった新大陸は比較的高いという傾向は読みとれる（図10）。

ところが、かなり地理的に違っている集団でも、遺伝子頻度が似かよっていることがある。これは遺伝的浮動により遺伝子頻度に収斂が生ずるからである。図11は遺伝的浮動をコンピュータ・シミュレーションで示したものである。縦軸が遺伝子頻度、横軸が世代である。遺伝子頻度0.5から出発して、遺伝子頻度がだんだんあがってくる場合もあれば、あるいは下がっていってしまう場合もいろいろある。

一般には、時代がたてばたつほど、最初同じだった集団が二つに枝分かれして別々の進化の道を歩み出し、遺伝子頻度が変わるということが期待される。ところが、なかにはかなり長い間たった後に、また

```
101  138  ABO: A, B, O
.1156  .0873  .7972
104  84  Fy: a, b & -  (Duffy)
.8457  .1543
105  40  Kell: k, K
.00  1.00
109  91  MN: M, N
.4286  .5714
111  74  P: P1, P2 & p
.2740  .7260
112  138  Rh: D, d
1.0  .0
113  138  Rh: CDE,CDe,cDE,cDe,CdE,Cde,cdE,cde
.0  .8555  .0362  .0889  .0  .0  .0  .0
120  74  Ge: a, b & a-
.7987  .2013
218  138  Hp: 1, 2  (Haptoglobin)
.6316  .3684
221  138  Pi: M, F (Alfal-antitrypsin)
.9964  .0036
223  138  Tf: C, D-Manus (transferrin)
.9819  .0181
307  138  ACP: a, b, c
.1775  .8225  .0
312  137  ADA: 1, 2
.9161  .0839
313  138  AK: 1, 2
1.0  .0
324  2  G6PD: +, -
.9788  .0212
335  138  LDHA: 1, Cal-1
1.0  .0
336  138  LDHB: 1, Cal-1
1.0  .0
342  138  PepA: 1, 2
1.0  .0
343  138  PepB: 1, 2, 3, 4
1.0  .0  .0  .0
347  138  PGM1: 1, 2
.9016  .0984
349  138  PGM2: 1, 2
1.0  .0
351  138  6PGD: A, C
.8152  .1848
353  138  PGK: 1, 2, 4
1.0  .0  .0
Manus Islanders -- Malcolm et al. (1972) HH 22:305-322
```

図8．集団ファイルの例（マヌス島集団）

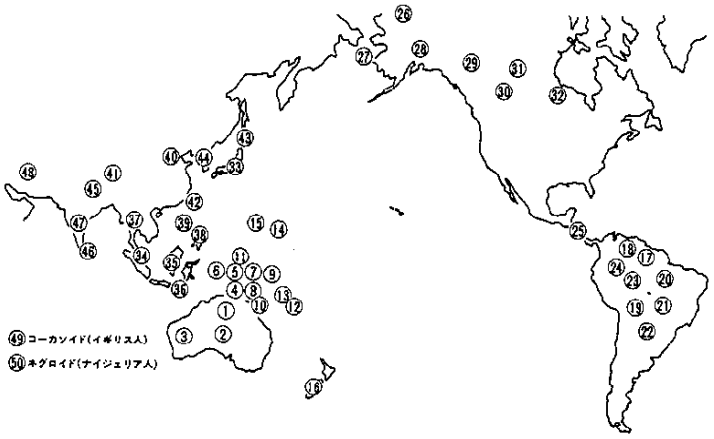

図9．比較した50集団の地理的位置
①～⑯：オセアニア；⑰～㉕：中央アメリカ及び南アメリカ；㉖～㉜：北アメリカ；㉝～㊺：東アジア及び東南アジア：㊻～㊿：その他。

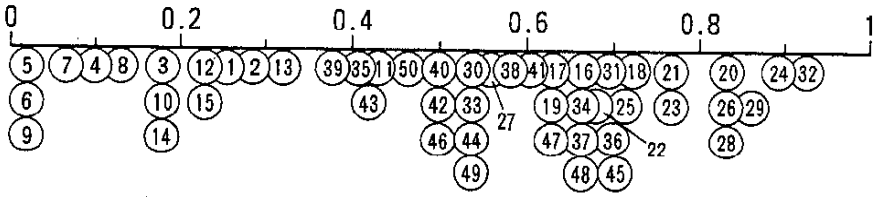

図10. MN血液型M対立遺伝子の頻度分布（以下図14まで、番号は図9のものに対応する）

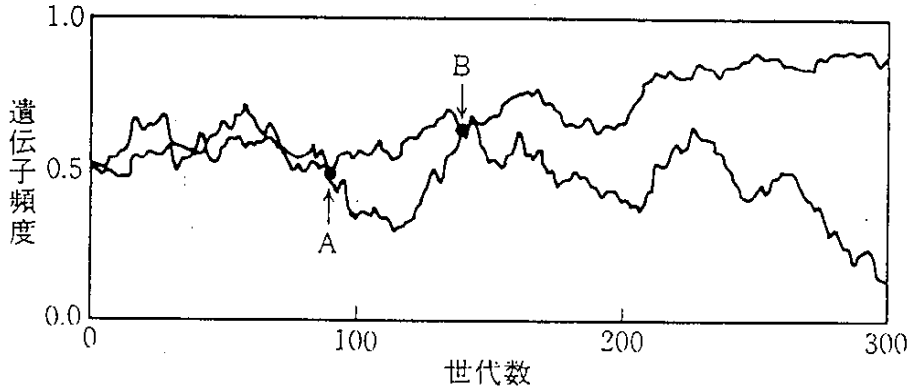

図11. 遺伝的浮動のコンピュータシミュレーション

同じ遺伝子頻度になることがある（グラフが交差している点A、B）。つまり、たまたまこの時刻に調べると2集団の遺伝子頻度が同じになるということがあるわけである。したがって一つの遺伝子座だけ調べて、遺伝子頻度が同じだったから二つの集団はつい最近分化したという議論は危険であるということは、これから明瞭に示すことができると思う。このため、たくさんの遺伝子座を使う必要があるのである。

用いる遺伝子座の数が遺伝的近縁図に与える効果を以下に示す。先ほどの50集団のABOの遺伝子座だけを使って遺伝距離を計算し、それで遺伝的近縁図を作成すると、中南米、北米、オセアニア、ノンモンゴロイドというようにある程度のクラスターはできるが、地理的にかなり散らばっている(図12)。遠い昔に分岐した集団であっても、遺伝子頻度がたまたま似てくると、遺伝距離も近くなるので、このようなクラスタリングが起こるわけである。

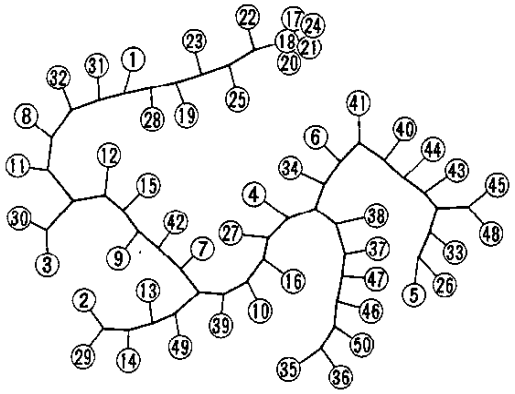

図12. ABO血液型データのみを用いて得られた50人類集団の遺伝的近縁図

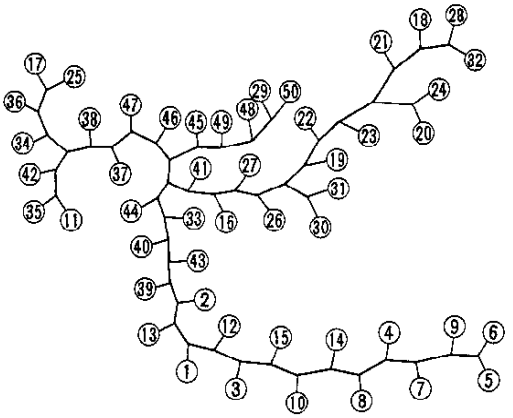

図13. ABO, Rh, MN血液型データを用いて得られた50人類集団の遺伝的近縁図

遺伝子座の数を3座位（ABOとRhとMN）に増やしてみよう（図13）。すると新大陸の集団が大体クラスターしてくる。それでもまだアウトライヤーがぽつぽつとある。たとえばアフリカの集団と北米の集団がクラスターしてしまう。こういうことが起こるわけである。

6座位にすると、さらに集団の地理的近縁関係を反映するようになる（図14）。このように集団の数を増やしていくと、先ほどからも問題になっているように、わずか6座位といっても世界中の50集団全部で調べられているわけではない。残念ながら集団の数を39個に減らさざるをえなかった。

最終的に30個の集団で12遺伝子座位をしらべた。図15はこれらのデータから遺伝距離を計算して近隣結合法で遺伝的近縁図を描いたものである。すると、今度はようやく北アメリカの集団（全部原住民）、南アメリカのインディアン、アジアの集団、オセアニアの集団（サモア人）がクラスターをなす。サモア人はポリネシア人なので、地理的にはオセアニアであるが、われわれアジアの人間と似ている。あとは非モンゴロイドがきれいに外に出て、すなわち枝Aでモンゴロイドと非モンゴロイドの分岐が示されている。このうちインド人がモンゴロイドに一番近くなっているが、これはインド・ビルマからシベリアのほうにいくラインで、まさにモンゴロイドと非モンゴロイドが分かれるわけである。このように遺伝子座の数を増やして、ある程度これならまあまあ合理的であろうかというものが初めて示されるわけである。

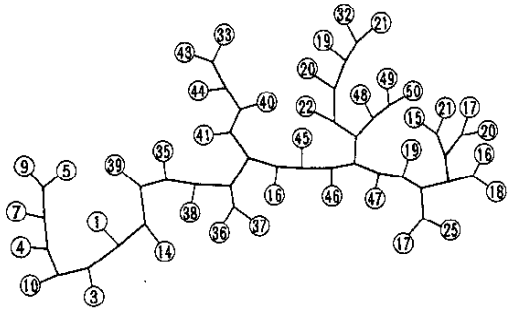

図14. 6遺伝子座のデータを用いて得られた39人類集団の遺伝的近縁図

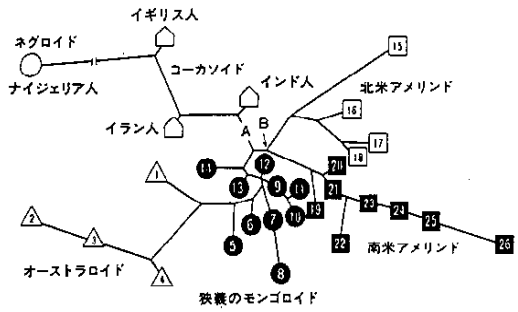

図15. 12遺伝子座のデータを用いて得られた30人類集団の遺伝的近縁図

そこでおもしろいのは、まだ暫定的な結果ではあるが、南アメリカと北アメリカの集団がかなり違う点である。ただ北アメリカといっても、これは残念ながらアメリカ合衆国のインディアンのデータは少なくて、たいていはカナダの集団である。いずれにせよ南北アメリカの分岐が非常に深いようだというところがおもしろいと思う。とくに、図15にオーストラリア原住民やパプアニューギニアのデータが入っているわけだが、それはアジアから枝分かれしている。それよりも前にアメリンドがアジアから分かれているように見える。それが本当かどうかはまだわからないが、そうすると1万年～2万年前どころか、3万年～4万年前という古い時代に旧大陸と新大陸の集団が分岐してもよいような気がする。実際、最近言語学のほうでアメリカのバークレーの研究者ジョアンナ・ニコルズが、オーストラリア原住民の言語の多様性とアメリカインディアンの言語の多様性は同じくらいである、だから新旧大陸の分岐年代もかなり古いのではないか、と言っている。しかし、将来、もっとデータを増やしてみなければ、確実なことは現在の時点では言えないだろう。

*　　*　　*

本稿は、重点領域研究「先史モンゴロイド集団の拡散と適応戦略」公開シンポジウム（1992年1月13日・14日・15日　於東京大学）のトピックIII「モンゴロイド諸集団の起源・系統」における研究発表である。

参考文献

Roychoudhury, A. and M. Nei (1988) Human Polymorphic Genes : World Distribution. Oxford University Press, Oxford.

斎藤成也（1992）人類遺伝子の系図．科学朝日，52(4)：48-52.

斎藤成也（1992）アメリカ大陸への人類の移動と拡散．赤澤威・坂口豊・冨田幸光・山本紀夫(編)アメリカ大陸の自然誌第2巻―最初のアメリカ人，pp.57-103．岩波書店．