MEASURE AND REPRESENTATION OF THE GENETIC SIMILARITY BETWEEN POPULATIONS BY THE PERCENTAGE OF ISOACTIVE GENES

Alicia SANCHEZ-MAZAS[1], Laurent EXCOFFIER[1]
and André LANGANEY[1]

ABSTRACT

A similarity index allowing comparisons of human populations has been defined as the common "Percentage of Isoactive Genes" or PIG, which can be calculated from any gene frequency distribution characterizing two populations. The complement to one of this value has been proved to be a distance, a measure which can be used in most techniques of cluster analysis as well as in usual representations of multivariated data (dendrograms, etc..). Furthermore, the formula can be generalized to a set of populations. From a biological point of vue, PIG values are particularly meaningful, whereas other previously defined indices which tend to measure similarity or difference between populations neither have such an advantage, nor can they be expressed by a single and clear number like a percentage. Moreover, comparisons using PIG indices depend at first on the fluctuations of the most frequent genes of a distribution; on the other hand, different measures principally reflect the variations of low frequency genes, which unfortunately are nearly always poorly estimated, due to weak samples of populations.

Interestingly, PIG values can be applied to any frequency distribution of any kind of objects. They can also extend to a whole set of distributions by using a mean value. When applying these measures to a series of polymorphic genetical systems such as ABO, Rhesus, MNSs, Gm and HLA, one can account for the variability or the homogeneity particular to the different human groups, with simple and objective criteria.

INTRODUCTION

Practical considerations on data of human genic frequencies lead us to restart work on their representation, a subject that many authors have tackled in three very distinct perspectives :

-The measure of kinship based on a genetic model [12], [9], [10], [5].

-The measure of genetic distance according to a phylogenetic model [1].

* This study was supported by the Fonds National Suisse de la Recherche Scientifique, Grant Nº 3.954-0.84.

THEORIA - Segunda Época - Año II
Curso 1986-1987, nº 4, pp. 143-154

-The measure of arbitrary distances for comparative purposes [8], [4].

The first two types of measures assume constrictive models, which are practically never found in existing human populations. Their use was quite successful since the estimators of kinship coefficients or phylogenetic distances had a structure with good empirical distance indication [6]. The angular distances of the second type model present the handicap of being very sensitive to frequency fluctuations of the rarest genes, which, in concrete cases, are always poorly estimated [7], [8].

Distances or similarity indices proposed by Jacquard [4] have no biological sense and may only be used for representations or visualisations of population differences.

It would be possible to make a measure of similitude between populations based on any distance or empirical index system computed from genic frequency vectors, and this for different polymorphic markers known on this set of populations. But it appeared to us easier and more significant to compute, for each couple of populations, the common Percentage of Isoactive Genes (PIG) on a system or the average PIG on a set of systems. This avoids the delicate problem of system ponderation, especially when genetic systems exhibit a different number of alleles. By the way, such indices will favour differences or similarities due to the most frequent genes whose frequencies are better estimated, contrarily to angular distances.

METHOD

1) *Distance description*

We are using as a similarity index the Percentage of Isoactive Genes (PIG) shared by two or more populations and as a distance its complement to one, already evoked by Gregorius [3]. In fact, this method is very general and can be applied far beyond genetics (i.e. to population comparisons but also to any object characterized by one or more frequency distribution). It allows clustering tests on populations, as most cluster and multivariate analysis methods.

It has also the advantage of estimating the similarity or difference between populations either with a pourcentage or with a shared gene frequency, which are concrete values with evident meaning contrarily to arbitrary similarity indices or distances commonly used.

This leads us to define, for any number of concerned populations and for any frequency distribution established on this set of populations the common Percentage of Isoactive Genes by

$$\mathrm{PIG}_{ij} = 100 \sum_{s=1}^{k} \min\, (f_{is}, f_{js})$$

where k is the size of the frequency vector, f_{is} the s^{th} frequency of the i^{th} population and PIG_{ij} the pourcentage of isoactive genes shared by populations i and j.

One verifies easily that

$$D_{ij} = 1 - (\mathrm{PIG}_{ij}/100) = 1 - \tfrac{1}{2} \sum_{s=1}^{k} [f_{is} + f_{js} - |f_{is} - f_{js}|]$$

and becomes

$$D_{ij} = \tfrac{1}{2} \sum_{s=1}^{k} |f_{is} - f_{js}|$$

In such a context, we shall call this new mathematical distance the *PIG distance.*

The PIG's may advantageously be defined on any number of populations, measuring the amount of similarity of a set of n populations , according to the equality

$$n\mathrm{PIG}(1,....,n) = 100 \sum_{s=1}^{k} \min\, (f_{1s},....,f_{ns})$$

We shall define as well

$$nD(1,....,n) = 1 - (n\mathrm{PIG}(1,....,n)/100)$$

which is not a mathematical distance, but may measure the maximal range of dissimilarity.

One will be able, at first sight, to recognize structures in a frequency distribution when comparing the nPIG's of high level with 2PIG's of level 2. Groups having the same structure (i.e. samples of a single population) diverging only by random fluctuations will have nPIG's with high values, approximately of the same order as *2*PIG's. On the other hand, in a gradient of populations, *n*PIG's may be very low though *2*PIG's may have high values between neighbouring populations. So there is a possibility of representing and discussing, with only one type of clearly understandable index, what one usually characterizes by many types of arbitrary distances (euclidian, ultrametric, etc...).

2) *Clustering technique*

For any set of populations, any complete and complementary frequency distribution can be used in order to compute a positive symmetrical matrix M_{Dij} made up of *2*PIG's. Usual techniques of factorial analysis allow of course the visualisation of the PIG distances by their projections on a series of orthogonal axes in a k-D space. As well, classical tree construction methods (minimal, maximal, average linkage and centroid cluster) may use D*ij* as distances in order to cluster populations with additional precision as for the groups obtained. As a matter of fact, it is possible to associate to each node of the tree the *n*PIG value common to the clustered populations. Thus a high *n*PIG for a cluster will show its homogeneity. On the other hand, a low *n*PIG will, when *2*PIG's are high, reflect a gradient-like structure of the cluster.

These indices computed on two or more populations lead us to define a new clustering method whose graphical output is a dendrogram as well. However it is not only based on the comparison of D_{ij}'s taken two by two, but also at higher levels on the nD($1,......,n$) indices common to a set of populations. At each clustering level, one will choose as a criterion min(nD_{gh}) with

$$nD_{GH} = 1 - (nPIG_{GH}/100) \ ,$$

nPIG$_{GH}$ being the index of similarity computed on a set of n populations which belong to the *G* and the *H* clusters taken as a whole. Thus the outcoming tree has been built with objective clustering criteria, without computation of fictive "mean populations" as in the average linkage. The stability of the final structure of the dendrogram will be evaluated by comparison with usual clustering techniques.

3) *System of permutations*

As to avoid an arbitrary order for the populations on the vertical axis, it is convenient to use a system of permutations in the final tree. This system, without disturbing the structure of the dendrogram, simply rotates its branches. It arranges the populations by minimizing the sum of distances between adjacent populations on the vertical axis. In practice, it's a matter of :

-either placing a unique population k on one side or the other of a cluster (i,j) of two populations,

if $D_{ik} < D_{jk}$, then the order is : kij,
if $D_{ik} > D_{jk}$, then the order is : ijk,

-or arranging the populations which belong to two clusters (i,j) and (k,l), thus the order will be $ijkl$ if $D_{jk} = \min(D_{ik},D_{jk},D_{il},D_{jl})$,

-and finally for clusters with more than two populations, one will generalize the method by comparing the D_{ij}'s of the flanking populations of the already arranged clusters.

3)*Generalization to a set of frequency distributions*

As we have already done for a certain number of different genetic systems taken one by one, we can also define a mean similarity or a mean dissimilarity index on all systems.

If we call PIG_{xij}, the PIG corresponding to the x^{th} component of the t frequency distributions, then

$$\overline{PIG}_{ij} = {}^{1}/_{t} \sum_{x=1}^{t} PIG_{xij} \quad \text{and} \quad \overline{D}_{ij} = 1 - (\overline{PIG}_{ij}/100)$$

In the same way, we may define $n\overline{PIG}(1,....,n)$ on the totality of the systems and construct a "mean dendrogram" of populations which will synthesize the information gathered by the different genetic polymorphisms studied so far.

APPLICATION

As a first example, we have calculated the PIG matrix of 15 populations out of the genic frequency distributions of the Rhesus system (Table 1). Data collected from publications have been checked and thus are reliable [11]. The eight haplotypes considered here were defined by the antigens DCcEe. Figure 1 shows a linear arrangement of these populations, calculated with an algorithm developped by G.M. Lathrop[2], using as a distance the complement to one of PIG. We have written down PIG values of neighbouring populations. On the horizontal axis, we have plotted gene frequencies.

2	3	4	5	6	7	8	9	10	11	12	13	14	15	Populations:
88	68	58	32	33	33	14	12	11	13	15	13	11	12	1: Sandawe
	72	61	42	41	45	23	23	20	22	24	19	20	20	2: Yoruba
		78	50	48	52	23	25	23	23	24	23	20	22	3: Tutsi
			70	67	71	44	43	41	44	45	44	41	42	4: Arabs
				92	94	72	72	69	70	70	67	70	65	5: Plati
					89	74	75	72	73	73	73	71	63	6: Arora
						69	70	66	68	68	63	66	67	7: Dari
							95	94	86	89	73	78	82	8: Koreans
								91	84	87	71	77	79	9: Japanese
									87	90	73	79	86	10: Papago
										95	87	90	76	11: Chinese
											82	86	79	12: Polynesians
												94	62	13: Indonesians
													68	14: Melanesians
														15: Quechua

Table 1: 2PIG matrix of 15 populations of the world, computed from their rhesus gene frequencies (the haplotypes are defined by the antigens D C c E e).

Such a representation allows us to visualise immediately genetic differences and similarities among populations. At the same time, it gives us a concrete measure of the latter. On figure 1 for example, three principal human groups are visible : the first one seems to cluster the whole of "oriental" populations (from Asia, Oceania and America), the second one gathers "occidental" populations (from Europe, India and Near East), and a last one consists mainly of black people from Africa, with Tutsi being however somewhat different. Inside each of these groups, PIG values reach 90 % or even more; on the contrary, each cluster share approximatively 70 % of its genes with each other. Arabs show intermediate frequencies between African Blacks and white people, though they are closer to the latter. The two Basque samples (from Spain and France) are very similar, with exceptional high frequencies of the r gene, never found anywhere else in the world. At the same time, R1 frequencies are lower than in the other populations of the occidental group, yet the general distribution of Basque genes resemble those of Europeans.

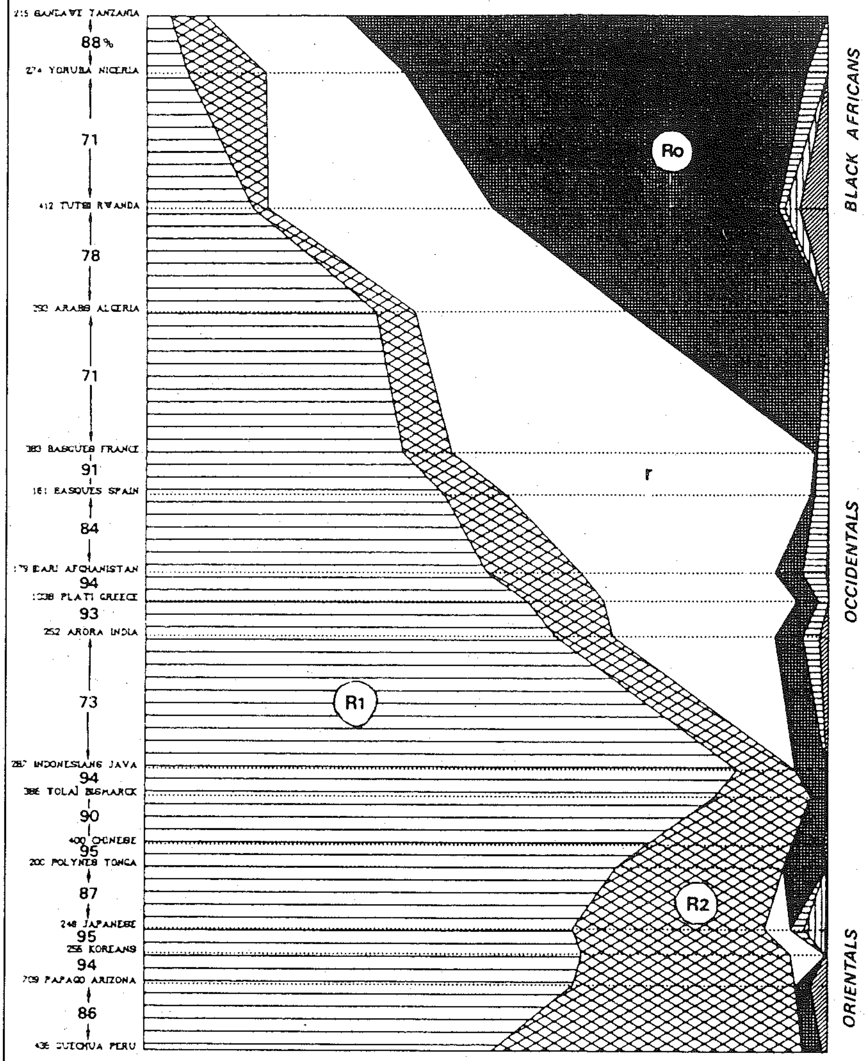

Figure 1: Rhesus frequencies in the world. 2PIG values between adjacent populations are indicated. G.M. Lathrop's permutation algorithm uses 1-PIG/100 as a distance.

As a second example, we have built mean dendrograms from a series of genetic systems, which are :

- three erythrocitary systems : ABO, MNSs and Rhesus,
- one seric system : Gm,
- one histocompatibility system : HL-A.

As a matter of fact, we found it fundamental to calculate PIG values not only from frequencies characterizing a single system but several: we shall be able to compare results obtained with each of them and the whole, in order to confirm or infirm our hypotheses on population kinships.

Two clustering methods have been used : centroid (figure 2) and pig cluster (figure 3). In this way, we obtain mean dendrograms for 14 populations. At each clustering level of the trees, we have mentioned the mean value of *n*PIG of the *n* gathered populations. The arrangement of the populations on the vertical axis is obtained by permutations (method described earlier in this paper).

Both representations show similar structure in the principal computed clusters : as in the first example, we can distinguish a black african group, an occidental and an oriental one. Inside each of them , however, and particularly in the latter, the structure is unstable: in the first case, Amerindians appear to be the last population to be grouped with the oriental ones whereas in the second picture it seems as if Micronesians and Australian Aborigines formed a distinct group. Nevertheless, the methods employed here allow us to use any number of populations as well as any number of genetic systems. Thus we can incorporate at any time more detailed information, that enables us to interpret stable or unstable dendrogram structures like here, and discuss them in a population kinship problematic. These examples show that a large investigation using many systems reaches an appreciable degree of reliability.

DISCUSSION

The interest of PIG's is to give a very meaningful measure of genetic similitude in a set of populations for one or more systems and without unchecked hypotheses on the mechanisms leading to the observed dissimilarity. The PIG variance of two populations on a set of systems may be a good indicator of possible differenciating

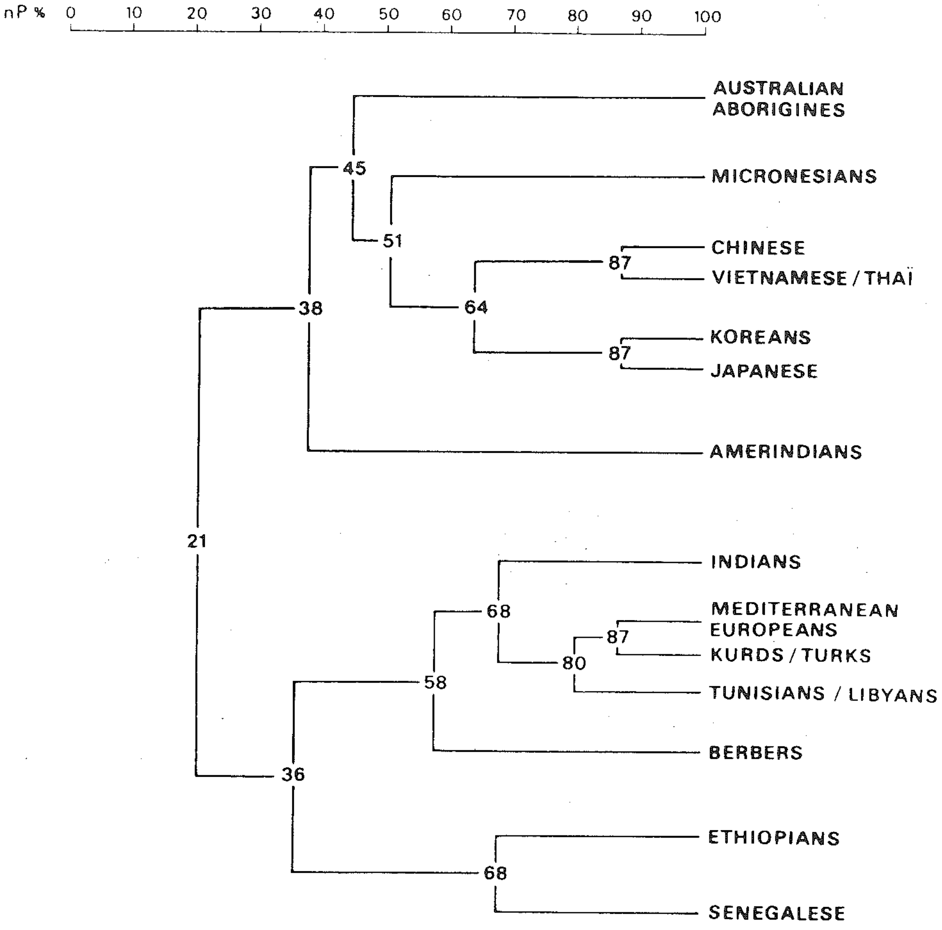

Figure 2: Mean centroid cluster dendrogram of 14 populations of the world, obtained from the genic frequency distributions of the systems ABO, Rhesus, MNSs, Gm and HLA (loci A and B).

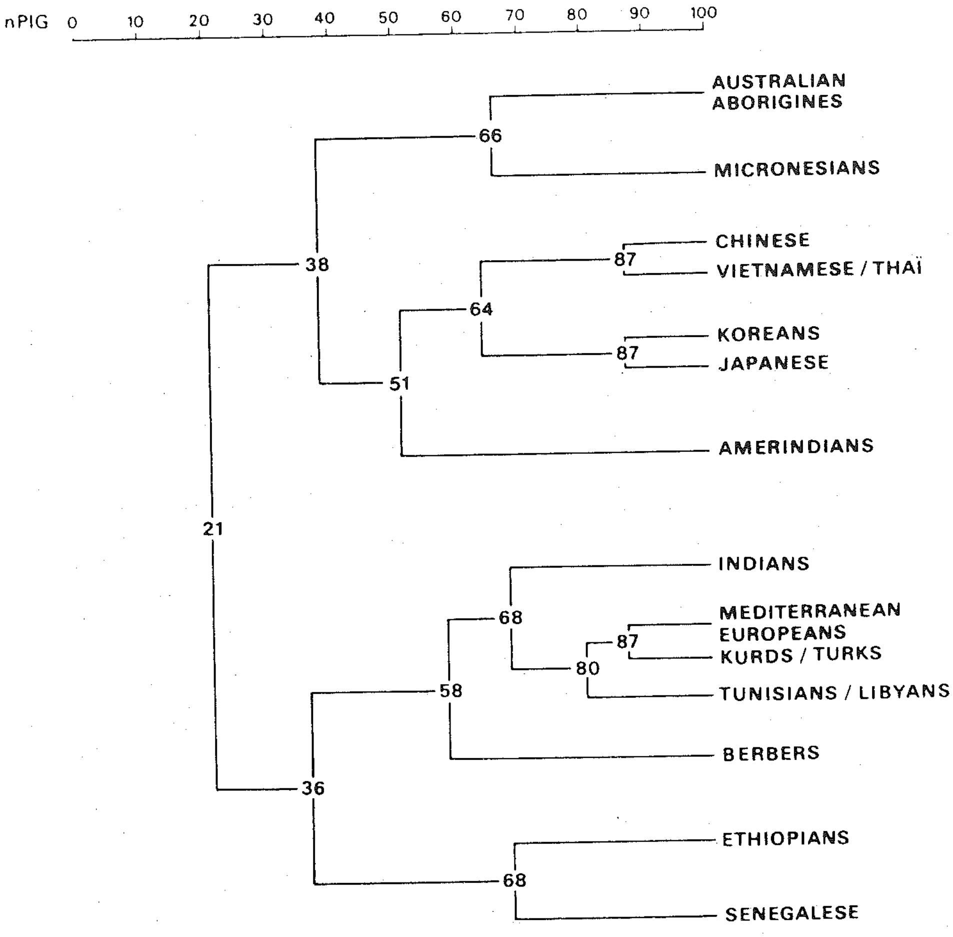

Figure 3: Mean PIG cluster dendrogram of 14 populations of the world, obtained from the genic frequency distributions of the systems ABO, Rhesus, MNSs, Gm and HLA (loci A and B).

mechanisms. On the other hand, a low variance and high PIG's will be a valuable presumption of a common and recent descent with a late separation.

In the case of a population having high PIG's and low variance with several other populations, it will be possible, with a reasonable incertainty, to foretell the genic frequencies of a system studied in the other populations after having used multiple regression techniques on the PIG's. This would avoid time consuming and technically expensive systematic research of certain genetic systems. This would also allow us to be on the eve of practical applications in genetic epidemiology (*a priori* estimations of genetic pathology incidence).

CONCLUSION

PIG, *n*PIG and $\overline{nPIG}$ are without doubt the most simple, reliable and meaningful measures of the genetic differentiation between populations. They are free of uncontrollable theoretical hypotheses and over-simple models in population genetics. They give an empirical measure of the common genetic heritage of two or more populations. They allow discussion on some of the causes of their differenciation thanks to non-arbitrary scaled graphical output and clear visualisation of distances.

Recent data in geographical haematology are yet neither numerous enough nor reliable enough for this method to be systematically applied, but it shows promises for epidemiological applications and Health policy.

FOOTNOTES

(1) Laboratoire de Génétique et Biométrie, Université de Genève, 12 rue Gustave-Revilliod, CH-1227 GENEVE, SWITZERLAND/SUIZA.

(2) Personnal communication described in [11].

LITERATURE

[1] CAVALLI-SFORZA L.L. and EDWARDS A.W.F., *Amer. J. Hum. Genet.*, 19, nº3, 1967, p.234.

[2] GILLOIS M., Thesis, University of Paris, 1964.

[3] GREGORIUS H.-R., *Math. Biosc.*, 41, 1978, p.253.

[4] JACQUARD A., *Cah. Anthropol. Ecol. Hum.*, 1, nº1, 11.

[5] LALOUEL J.M., Thesis, University of Paris 6, 1975.

[6] LANGANEY A., *Cah. Anthropol. Ecol. Hum.*, 2, nº1, p.11.

[7] LANGANEY A., *Population*, 6, 1979, p.985.

[8] LANGANEY A. and LE BRAS H., *Population*, 1, 1972, p.83.

[9] MALECOT G., Masson et Cie, Paris, 1948.

[10] MORTON N.E., YEE S., HARRIS D.E. and LEW R., *Theor. Pop. Biol.* 2, 1972, p.507.

[11] SANCHEZ-MAZAS A., Diplôme, University of Geneva, 1985.

[12] WRIGHT S., *Genetics*, 28, 1943, p.114.