Indian J. pure appl. Math., 26(6) : 503-519, June 1995

SOME STATISTICAL MODELS, METHODS AND APPLICATIONS IN GENETICS

PARTHA P. MAJUMDER

Anthropometry & Human Genetics Unit, Indian Statistical Institute, Calcutta 700 035

Some statistical models and methods used in genetical studies of populations are described. In particular, estimation procedures pertaining to parameters of models used for measuring genetic structures of contemporary populations, within- and between-population genetic diversities and population phylogenies are discussed.

A second section of this paper describes models and procedures pertaining to genetical studies in families. Models for estimating genetic and environmental components in the determination of a quantitative character are described. An application to family data on blood pressures is presented. A newly developed multilocus epistatic model is described, likelihood of non-randomly sampled family data is derived and an application to understanding the genetics of a pigmentary disorder (vitiligo) is presented.

INTRODUCTION

Statistics has always occupied a central place in genetical investigations. Many of the commonly used statistical methods (e.g., regression) were, in fact, developed in the course of genetical investigations. Because the foundations of Mendelian genetics rest on a set of probability laws, the use of statistical models and methods is natural to the analysis of genetic data. Contemporary statistical models in genetics remain to be crude approximations of reality because of the intrinsic complexity of biological systems. However, refinements continue to be made. Statistical applications in genetics range from providing statistical models of Darwinian evolution to the determination of the probability that an individual accused of having committed a murder is indeed the culprit. In between are a whole host of problems in which statistical models and methods play key roles. These include assessment of genetic relationships among contemporary populations, determination of mode of inheritance of a character, estimation of mutation rates due to various factors such as ionizing radiation, identification of breeds of commercial importance and designing efficient breeding strategies, assessment of relative roles of genes and environment in the determination of a character, mapping genes controlling a genetic character, etc. Since it is infeasible, in a single article, to describe the myriad of statistical models and methods used in this vast territory, we shall restrict our attention to only some specific areas. Further, this article will focus primarily on models and applications to humans. It must, however, be emphasized that many of the statistical models and

methods discussed below are applicable, and have actually been applied, to non-human species also. To avoid delving into specialized biological concepts and nomenclatures, we shall primarily discuss simple models and applications.

Genetic Studies in Contemporary Populations

The major concerns of genetical studies of contemporary polulations are : (i) estimation of genetic diversity within a population, (ii) measurement of genetic distance between a pair of populations, (iii) study of genetic relationships among a set of populations, and (iv) investigation of the genetical structure of a set of subdivided populations. For such studies, the data comprise maximum likelihood estimates (m.l.e.'s), x_{alp}, of allelic proportions, π_{alp}, where π_{alp} denotes the proportion of the ath allele ($a = 1, 2, ..., A_L$) at the lth genetic locus ($l = 1, 2, ..., L$) in the pth population ($p = 1, 2, ..., P$). Obviously, $0 \leq \pi_{alp} \leq 1$, and $\Sigma_{a=1}^{A_L} \pi_{alp} = 1$. In genetical literature, the allelic proportions, π's, are called allele frequencies, which will henceforth be used.

Gene Diversity

The extent of variation of allele frequencies at a genetic locus within a population is termed as gene diversity. An average of gene diversities over several loci is termed as the average gene diversity. While it is possible to use any of the standard entropy measures, such as the Shannon information measure, for measuring gene diversity at a locus within a population, the most commonly used measure is the Gini-Simpson index, which affords a biological explanation (viz., the probability that two randomly chosen genes are not identical) and has been popularized in genetics by Nei[18]. For simplicity, if we consider a single population and a single genetic locus, the gene diversity is defined as :

$$h(\boldsymbol{\pi}) = 1 - \Sigma_{a=1}^{A} \pi_a^2. \qquad \ldots (1)$$

(We have dropped the suffixes l and p, since they are irrelevant now.) This is a measure of the proportion of heterozygous individuals in a randomly mating population. Obviously, this measure has many desirable properties : it yields the same value for any permutation of $(\pi_1, \pi_2, ..., \pi_A)$; is ≥ 0 (= 0 iff $\pi_i = 1$ for some $i = 1, 2, ..., A$, and $\pi_j = 0$ for all $j \neq i$); and attains its maximum value for $\boldsymbol{\pi} = \boldsymbol{\varepsilon} = (1/A, 1/A, ..., 1/A)$. We also note that the maximum value of $h(\boldsymbol{\pi}) \to 1$ as $A \to \infty$. Are there other measures satisfying these properties ? Rao[27] proved that any measure $h(\boldsymbol{\pi})$ satisfying the above properties, the property $h\{(\boldsymbol{\pi} + \boldsymbol{\varepsilon})/2\} = c.\{h(\boldsymbol{\varepsilon}) - h(\boldsymbol{\pi})\}$ (where c is a constant) and a regularity condition, must be of the form

$$h(\boldsymbol{\pi}) = c_1 \left[1 - \Sigma_{a=1}^{A} \pi_a^2 \right] + c_2,$$

where $c_1 > 0$ and c_2 are constants.

The natural estimator of $h(\boldsymbol{\pi})$ is

$$h(\boldsymbol{\pi}) = 1 - \Sigma_{a=1}^{A} x_{a\cdot}^2 \qquad \ldots (2)$$

Now, $E[h(\boldsymbol{\pi})] = \{1 - (1/n)\}.(1 - \Sigma_{a=1}^{A} \pi_a^2)$, where n denotes the sample size. Hence, $h(\boldsymbol{\pi})$ is only asymptotically unbiased. However, in practice the bias is quite small since in most studies $n \approx 100$. The variance of $h(\boldsymbol{\pi})$ is (Nei and Roychoudhury[22])

$$V[h(\boldsymbol{\pi})] = [2(n-2)\{\Sigma x_i^2 - (\Sigma x_i^2)2\} + \Sigma x_i^2 - (\Sigma x_i^2)^2]/[n(n-1)]. \ldots (3)$$

The unbiased estimator of $h(\boldsymbol{\pi})$ is : $h(\boldsymbol{\pi}) = \{n/(n-1)\}.(1 - \Sigma_{a=1}^{A} x_i^2)$. Asymptotic distributions of these estimators have been derived by Nayak[17], which can be used for performing tests of some hypotheses pertaining to gene diversity.

Genetic Distance

For the purpose of studying genetic variation between populations, a measure of genetic distance was first proposed by Sanghvi[23]. For a single locus, Sanghvi's distance is defined as :

$$G^2 = \sum_{a=1}^{A} (\pi_{as} - \pi_{at})^2/\{2(\pi_{as} + \pi_{at})\}, \qquad \ldots (4)$$

where s and t ($s, t = 1, 2, \ldots, P$) refer to two populations. (We have suppressed the suffix l, since we are considering a single locus.) It can be shown (see Chakraborty and Rao[5]) that this distance measure is the same as Mahalanobis' D^2, when the underlying distribution is multinomial instead of normal. This is because of the structure of the dispersion matrix of allele frequencies, which for a set of allele frequencies $(\pi_1, \pi_2, \ldots, \pi_{A-1}, \pi_A = 1 - \Sigma_{a=1}^{A-1} \pi_a)$ has as the elements : $\mathrm{Var}(\pi_i) = \pi_i(1-\pi_i)$ and $\mathrm{Cov}(\pi_i, \pi_j) = -\pi_i \pi_j$; $i \neq j = 1, 2, \ldots, A-1$. The inverse of the dispersion matrix has $1/\pi_i$ as the diagonal entries, and $(1/\pi_i + 1/\pi_j)$ as the off-diagonal entries. By using the pooled allele frequencies of the two populations s and t to compute the pooled dispersion matrix, the proof that G^2 and D^2 are equal is straightforward. With its roots branching out from the basic formulation of a distance measure by Mahalanobis, various other genetic distance measures have subsequently been proposed. We shall not provide a review of these measures here; the interested reader may consult Nei[20]. We shall, however, consider one other distance measure because of its biological interpretability and its widespread use. This measure is due to Nei[18], and is defined as

$$D = -\ln\left[\Sigma_{a=1}^{A} \pi_{as}\pi_{at}/\left\{(\Sigma_{a=1}^{A} \pi_{as}^2)\cdot(\Sigma_{a=1}^{A} \pi_{at}^2)\right\}^{1/2}\right]. \qquad \ldots (5)$$

The numerator of the term within brackets in the definition of D denotes the proportion of identical genes between two randomly chosen genomes, one drawn from population s and the other from population t. The denominator is a normalizing factor reflecting the proportion of genes that are identical between two randomly chosen genes from within each of the two populations. When many loci are considered, the terms in the numerator and the denominator are replaced by the averages of each of the terms taken over loci; for example, the numerator is replaced by $L^{-1} \sum_{l=1}^{L} \sum_{a=1}^{A_l} \pi_{als} \pi_{alt}$.

It must be noted that D is not a metric because it does not satisfy the triangular inequality. However, in addition to its biological interpretability, it is useful for evolutionary studies because the expected value of this distance measure increases, under certain assumptions (stated below), linearly with evolutionary time of divergence/isolation of the two populations under consideration. Most other proposed measures of genetic distance which are metrics, including Sanghvi's G^2, do not satisfy this property. Under the assumptions that (i) the populations are in steady states under mutation-drift balance, (ii) the sizes of the populations remain constant over generations, (iii) the rate of gene substitution is constant across loci and generations, and (iv) every mutation gives rise to a new allele,

$$E(D) = 2\mu\tau, \qquad \ldots (6)$$

where μ denotes the rate of gene substitution, and τ denotes the time of divergence/isolation of the two populations. Thus, populations which diverged earlier are expected to have a greater genetic dissimilarity compared to populations which diverged later — a desirable property of a distance measure. Further, eqn. (6) can be used to estimate the evolutionary time of divergence if the substitution rate (μ) is known. Rao[27] showed that the transformation $\cos^{-1}(e^{-D})$ renders D to a proper metric. But this transformation destroys the useful relationship of the distance measure with evolutionary time.

A natural estimator of D is obtained by substituting m.l.e.'s x_{as} and x_{at} for π_{as} and π_{at}, respectively, in eqn. (5). However, this natural estimator is biased. The asymptotically unbiased estimator is[19]

$$\hat{D} = -\ln \{\sum_{a=1}^{A} x_{as} x_{at} / [\{2n_s \sum_{a=1}^{A} x_{as}^2 - 1)/(2n_s - 1)\} \times \{(2n_t \sum_{a=1}^{A} x_{at}^2 - 1)/(2n_t - 1)\}]^{1/2}\}, \qquad \ldots (7)$$

where n_s and n_t are, respectively, the sample sizes from populations s and t. The sampling distribution of $\hat{D}$ is unknown. The variance of the natural estimator is derived in Nei and Roychoudhury[22].

Genetic Relationships

The study of genetic relationships among populations is performed by estimating the matrix of pairwise genetic distances, and then performing a cluster analysis using this distance matrix. For contemporary populations with short evolutionary times of divergence/isolation, the most popular clustering algorithm is the average-linkage method, also known as the UPGMA method[30]. However, when the populations under

consideration have long divergence times; for example, when one is analyzing genetic data of different species or phyla; the UPGMA may not perform well in reconstructing genetic relationships, which involves estimating the topology of the phylogenetic tree and its branch lengths. This phenomenon has been observed in many computer simulation studies. The major reason for this is that rates of gene substitution are often variable over long periods of evolutionary time and across loci. Various other clustering algorithms have been proposed to deal with this problem (see Saitou[28] for a recent review).

Even when one is dealing with contemporary populations, there is a sampling variance associated with each estimate of pairwise genetic distance. Therefore, conceptually this will lead to a variance in node positions (distances at which OTU's successively join to form clusters) of a reconstructed tree. Nei *et al.*[23] have suggested a method of computing the variance of a node which is a weighted sum of the variances of and covariances between distances of populations which are included in the node under consideration (see also, Chakraborty[2]). Felsentein[9] has suggested a bootstrap approach to this problem. We are currently investigating another computer-intensive approach which not only provides sampling variances of the nodes, but also provides an estimate of the level of confidence of a particular tree topology. Our approach comprises generating many simulated distance matrices based on the estimated distance matrix, and reconstructing a phylogenetic tree for each of these generated matrices. For the estimate of the distance, $\hat{D}$, between a pair of populations and its variance, Var $(\hat{D})$, we draw a random number from $U[\hat{D} - \beta\sqrt{\text{Var}(\hat{D})}, \hat{D} + \beta\sqrt{\text{Var}(\hat{D})}]$, where $\beta > 0$ is a constant. When this is done for all the $P(P + 1)/2$ pairs of populations (symmetry of the distance matrix is assumed), we obtain a simulated distance matrix. We then apply a clustering algorithm to this matrix to obtain a phylogenetic tree. The procedure is repeated a large number of times, which yields a frequency distribution of different phylogenetic topologies, from which the most parsimonious topology can be chosen. This frequency distribution is a function of many parameters, including β. The variances associated with each node of this parsimonious topology are easily estimated from the simulated trees conforming to this topology. Our results based on analysis of actual gene frequency data from human populations, albeit preliminary, indicate that for contemporary populations often little confidence can be placed on the dendrogram depicting genetic relationships reconstructed from the observed distance matrix.

Genetical Structure

When a population is subdivided into a set of related subpopulations, it is of interest to investigate the genetical structure of these subdivided populations. An effect of subdivision is an increase in the proportion of homozygotes in the total (pooled) population, which results in departures of genotype proportions from those expected under Hardy-Weinberg equilibrium. This effect is observed even if the genotype proportions within each subpopulation do not differ significantly from Hardy-Weinberg expectations. If these subpopulations are random mating units but are not completely isolated, that is, if there is exchange of mates among the subpopulations, then Hardy-Weinberg equilibrium is slowly restored with the passage of time. Thus, deviations of genotype proportions from Hardy-Weinberg expectations

shed light on the genetical structure of a subdivided population. The genetical structure is studied in terms of three correlations between two gametes uniting to form an individual. In a randomly-mating population, there is no correlation between homologous genes of uniting gametes relative to the gene frequencies in the whole population. Upon division into primarily intra-mating subpopulations, correlation between uniting gametes is expected. The parameters for studying genetical structure are : F_{IT} = correlation between gametes that unite to produce the individuals (I) relative to the gametes of the total population (T), F_{IS} = correlation between uniting gametes averaged over all subpopulations relative to those of their own subdivision (S), and F_{ST} = correlation between two gametes drawn randomly from within a subdivision (S), relative to gametes of the total population (T). Wright[31] showed that

$$1 - F_{IT} = (1 - F_{IS}) \cdot (1 - F_{ST}). \quad \ldots (8)$$

With reference to the deviations of genotype frequencies from Hardy-Weinberg equilibrium for a single autosomal locus, these parameters can be defined as (Nei[32])

$$F_{IT} = (H_T - H_0)/H_T, \quad \ldots (9a)$$

$$F_{IS} = (H_S - H_0)/H_S, \quad \ldots (9b)$$

$$F_{ST} = (H_T - H_S)/H_T, \quad \ldots (9c)$$

where

$$H_0 = \Sigma_{p=1}^{P} \; w_P \, (1 - \Pi_{ap}), \quad \ldots (10a)$$

$$H_T = \Sigma_{a=1}^{A} \; \bar{\pi}_{a\cdot} \, (1 - \bar{\pi}_{a\cdot}), \quad \ldots (10b)$$

$$H_S = \Sigma_{p=1}^{P} \; w_p \, \Sigma_{a=1}^{A} \; \pi_{ap} \, (1 - \pi_{ap}), \quad \ldots (10c)$$

Π_{ap} = proportion of individuals who are homozygous for allele a (= 1, 2, ..., A) in population p (= 1, 2, ..., P), w_p = size of the pth subpopulation divided by the total population size, $\bar{\pi}_{a\cdot} = \Sigma_{p=1}^{P} w_p \, \pi_{ap}$. In genetic terms, H_0 denotes the proportion of heterozygotes in the total population, H_T denotes the proportion of heterozygotes expected under Hardy-Weinberg equilibrium ("heterozygosity") in the total population, and H_S denotes the average within-subpopulation heterozygosity. When data on many loci are used, averaging is also done over loci.

Assuming that, for all p = 1, 2, ..., P, w_p = $1/P$ (an assumption which is generally made in practice because of lack of reliable knowledge of subpopulation sizes), Nei and Chesser[21] provided approximate unbiased estimators of H_T, H_0 and H_S. The exact unbiased estimators have been derived by Chakraborty and Danker-Hopfe[4]. These are

$$H_T = \Sigma_{a=1}^{A} \, [\, \bar{x}_{a\cdot} \, (1 - \bar{x}_{a\cdot}) + P^{-2} \; \Sigma_{p=1}^{P} \; x_{ap} \, (1 - x_{ap})/(2n_p - 1)], \quad \ldots (11a)$$

$$H_0 = 1 - P^{-1} \, \Sigma_{p=1}^{P} \; \Sigma_{a=1}^{A} \; X_{ap}, \quad \ldots (11b)$$

$$H_S = P^{-1} \Sigma_{p=1}^{P} \; 2n_p \, (1 - \Sigma_{a=1}^{A} \; x_{ap}^2)/(2n_p - 1), \qquad \ldots (11c)$$

where $\bar{x}_{a.} = \Sigma_{p=1}^{P} \; x_{ap}/P$, $X_{ap} = n_p^{(aa)}/n_p$, $n_p^{(aa)}$ = number of individuals who are homozygous for the ath allele in the sample of size n_p from population p.

When these estimates of H_T, H_0 and H_S are plugged into eqns. (9), the resulting estimators are consistent, but not unbiased, for F_{IT}, F_{IS} and F_{ST} as defined by Wright[31]. Approximate sampling variances of these estimators can be derived by using the procedure given in Nei and Roychoudhury[22]. However, virtually nothing is known about the sampling distributions, even asymptotic, of these estimators, and therefore tests of hypotheses concerning these parameters cannot be performed. Some heuristic test criteria for certain hypotheses are given in Cockerham[6]. The relative values of the estimates of F_{IT}, F_{IS} and F_{ST} are generally used in a descriptive manner. F_{ST}, which is necessarily positive, when "large" indicates that there is a strong effect of subdivision; that is, there is considerable genetic differentiation among the subdivided populations. While usually positive, if $F_{IS} < 0$, then it indicates that there is systematic avoidance of consanguineous mating within subdivisions. If there is systematic subdivision, whether into random mating subpopulations ($F_{IS} = 0$ and $F_{IT} = F_{ST}$) or into inbred groups, F_{IT} is positive but it can be negative if there is little or no systematic subdivision and there is prevailing avoidance of consanguineous mating. The various statistical procedures employed in the analysis of genetical structure of subdivided populations have recently been reviewed by Chakraborty and Danker-Hopfe[4] and Chakraborty[3].

GENETIC STUDIES IN FAMILIES

One of the major goals of genetics is to understand the causes of aggregation of a character (qualitative or quantitative) in families. Familial aggregation of a character can arise because individuals in families share common environmental factors (such as, food) or because they share genes. Thus, any statistical model for the analysis of the causes of familial aggregation must start with estimating the relative roles of genes and environment. It must be emphasized that sometimes environmental factors may be transmissible in the same way that genes are. For example, parents often play a leading role in imparting attitudes and values to their children, although not all children may pick up similar attitudes and values, as is often the case. Thus the transmission of a learned behavioral character may mimic Mendelian transmission even though it is not genetic. Therefore, a character being transmitted from parents to offspring in accord with Mendelian probabilistic laws is not a sufficient condition that it is genetic. Statistical models must take this fact into account. The ultimate proof that a character is genetically controlled is arrived at not in a statistical laboratory, but in a genetical laboratory by actually cloning the gene(s) controlling the character. However, statistical models and methods can and do provide guidance to geneticists in determining gene cloning strategies so that cloning efforts are not reduced to searching for a needle in a haystack. In this section, we shall

describe some of these models and provide some applications of (i) estimating the relative roles of genetic and environmental factors in the determination of a character, and (ii) determining the mode of inheritance of a genetic character. Because of space constraints, we shall not be able to discuss the statistical methods for determining whether two characters cosegregate in families. A review of these procedures can be found in Ott[24] and more recently in Ott[25].

Path Model

During the last five years we have been involved in several studies pertaining to the determination of genetic and environmental contributions to blood pressure levels — systolic (SBP) and diastolic (DBP). The motivation for these studies is that although essential hypertension — defined by the W.H.O. as SBP > 160 mmHg or DBP > 95 mm Hg — is a major risk factor for cardiovascular diseases, even "mild" hypertension, conventionally in the range of 90 to 104 mmHg DBP is viewed as a major public health burden (Labarthe[33]). Therefore, the belief is that an understanding of how blood pressure is influenced by genetic and environmental factors is the key to understanding the role played by hypertension in cardiovascular complications. The most recent study undertaken by us was among the Marwaris residing in Calcutta. The reasons for the choice of Marwaris as a study group, pertinent study details and their epidemiological profiles relating to blood pressures and lipid levels have already been published (Majumder *et al.*[14]). Suffice it to say that the Marwaris have a high prevalence (17%) of hypertension, and their lipid profiles are "worse" than the Caucasians living in the United States (who have poor lipid profiles from the cardiovascular standpoint). To study the relative contributions of genes and environmental factors, we have performed a path analysis of the blood pressure data collected from members of 210 nuclear families. Path analysis, developed by Sewall Wright, is a method based on algebraic manipulation of standardized regression coefficients to explain linear relationships between variables. The path model used in the present study is presented in Fig. 1. As may be seen from the legend to this figure, there is a variable called "environmental index" (*I*). Since family environment in its totality is not directly measurable, an index was created as an estimate of the environment, separately for SBP and DBP. The index is assumed to be a measure of the family environment alone, and the genetic correlation between the index and the corresponding blood pressure variable is assumed to be zero. We note that the existence of any such correlation is expected to yield an underestimate of the genetic heritability (the variance in the character explained by genetic factors) and an overestimate of the environmental heritability. The index (*I*) was created by a stepwise regression procedure using the anthropometric variables height, weight, and skinfold thicknesses — biceps, triceps and subscapular. Blood pressure levels were also simultaneously adjusted for the effects of other concomitant variables such as age, gender, education, occupation, tobacco use, alcohol use, contraceptive use, steroid use, presence of tension and major disease. The adjusted levels were then standardized. Details of the procedure are given in Majumder *et al.*[14]. Path analyses were performed on standardized blood pressure variables. Under the assumption of multivariate normality and an intra-class correlation structure, maximum likelihood estimates of familial correlations were

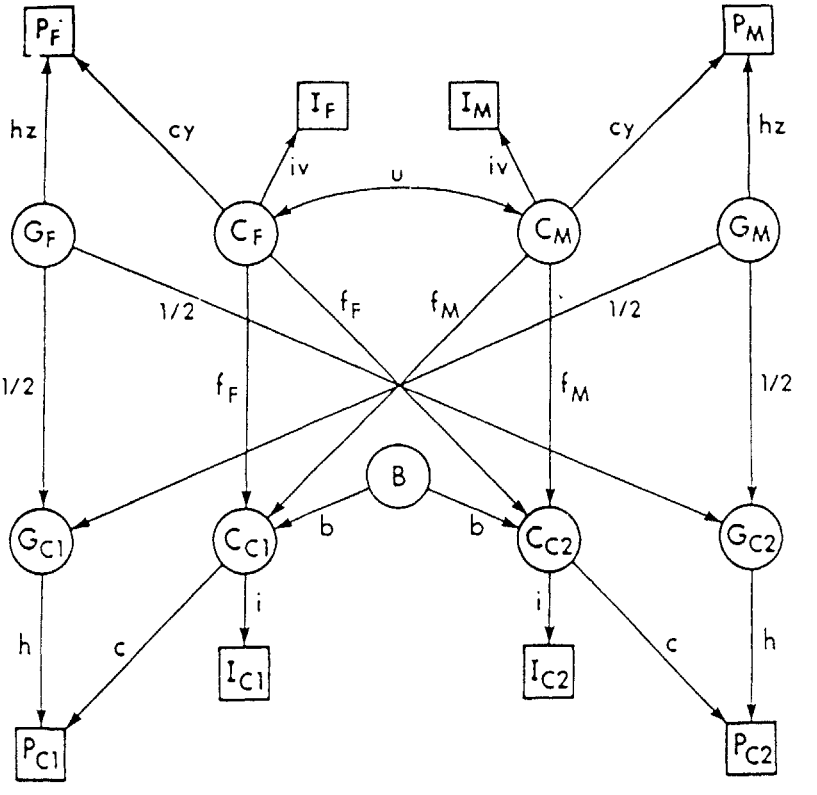

FIG.1 . Path diagram depicting biological and environmental (cultural) inheritance of blood pressure in nuclear families. P denotes phenotype (standardized blood pressure, systolic or diastolic), G denotes genotype, C denotes transmissible indexed (primarily cultural) effect, and B denotes nontransmitted common sibship environment. The subscripts F, M, C1 and C2 denote, respectively, father, mother and two children.

estimated, which are presented in Table I. The 10 parameters of the path model are defined in Table II. Marital resemblance, as measured by the correlation between environments of spouses (u) reflects the combined effects of social homogamy and cohabitation. Genetic and common environmental (cultural) heritabilities are h^2 and c^2, respectively, in children, and h^2z^2 and c^2y^2, respectively, in adults. Thus, possible intergenerational differences in heritabilities are taken into account in this model. It is pertinent to note that anthropometric variables have significant genetic components that may be correlated with plausible genetic factors involved in the determination of blood pressure levels. Therefore, the index I that we have created using anthropometric variables may contain some confounded genetic information in addition to environmental information. In view of this, it is appropriate to reinterpret the heritabilities. If h_t^2 and c_t^2 denote the true (unknown) genetic and cultural heritabilities, then $h^2 = (1 - \lambda)\, h_t^2$ and c^2 = indexed (combined) heritability. Parental environmental effects are also distinguished; effects of maternal and paternal environments on those of their children are denoted as f_M and f_F, respectively. In addition to a nontransmitted common sibship environment (B), separate indices are incorporated for each child.

TABLE I

Maximum likelihood estimates of familial correlations (r) and their standard errors (s.e.) for systolic (SBP) and diastolic (DBP) blood pressures

Variables	SBP		DBP	
	r	s.e.	*r*	s.e.
P_F, P_M	−.081	.106	.224	.106
I_F, I_M	.088	.110	.121	.112
P_F, I_M	−.049	.082	.147	.084
P_F, I_F	.362	.055	.457	.049
P_{C1}, P_{C2}	.324	.076	.252	.081
I_{C1}, I_{C2}	.463	.078	.352	.094
P_C, I_C	.491	.048	.470	.049
P_{C1}, I_{C2}	.315	.064	.259	.069
P_F, P_C	.134	.079	.177	.084
I_F, P_C	.142	.086	−.027	.096
P_F, I_C	.078	.091	.078	.093
I_F, I_C	.075	.102	−.006	.103
P_M, P_C	.245	.089	.202	.085
I_M, P_C	.154	.091	.121	.092
P_M, I_C	.125	.099	.193	.089
I_M, I_C	.105	.098	.108	.095

TABLE II

Definitions of parameters of the path model

Parameter	Definition
h	Effect of child's genotype on child's phenotype
hz	Effect of adult's genotype on adult's phenotype
c	Effect of child's environment on child's phenotype (= square root of cultural heritability)
cy	Effect of adult's environment on adult's phenotype
u	Correlation between parental environments
b	Effect of nontransmitted common sibship environment on child's environment
f_F	Effect of father's environment on his child's environment
f_M	Effect of mother's environment on her child's environment
i	Effect of child's environment on child's index
iv	Effect of adult's environment on adult's index

The phenotypes and indices of parents and children yield 16 distinct expected correlations by the use of standard rules of path analysis[10], which are presented in Table III. Maximum likelihood estimates of the path parameters were estimated using the methods given in Morton *et al.*[15] (see also, Rao *et al.*[26]). For identifying the most parsimonious model, various submodels which are nested within the general model

TABLE III

Expected correlations between variables observed on members of a nuclear family under the path model

Relation	Variables	Correlation
Martial	P_F, P_M	$uc^2 y^2$
	I_F, I_M	$ui^2 v^2$
	(P_F, I_M) or (P_M, I_F)	$ucyiv$
Parental	(P_F, I_F) or (P_M, I_M)	$cyiv$
Full sibs	P_{C1}, P_{C2}	$h^2/2 + c^2 \Psi$
	I_{C1}, I_{C2}	Ψi^2
	P_C, I_C	ci
	(P_{C1}, I_{C2}) or (P_{C2}, I_{C1})	$c \Psi i$
Parent-offspring	P_F, P_C	$h^2 z/2 + c^2 y(f_F + uf_M)$
	I_F, P_C	$civ(f_F + uf_M)$
	P_F, I_C	$cyi(f_F + uf_M)$
	I_F, I_C	$i^2 v(f_F + uf_M)$
	P_M, P_C	$h^2 z/2 + c^2 y (f_M + uf_F)$
	I_M, P_C	$civ(f_M + uf_F)$
	P_M, I_C	$cyi(f_M + uf_F)$
	I_M, I_C	$i^2 v (f_M + uf_F)$

Note : $\Psi = b^2 + f_F^2 + f_M^2 + 2uf_F f_M$

are tested. The various submodels tested in this study are : (i) no intergenerational differences in heritabilities ($y = z = 1$); (ii) no effect of assortative mating and cohabitation ($u = 0$); (iii) no extra sibling environmental effect ($b = 0$); (iv) no specific maternal environmental effect ($f_F = f_M$); (v) no genetic inheritance ($h = z = 0$); and, (vi) no cultural inheritance ($c = y = 0$, $i = v = 1$). Likelihood ratio test is used for examining the fit of a submodel in comparison with the general model. The fit of the general model is calculated by comparing the ln-likelihoods obtained for estimating the correlations with that obtained under the general model.

The maximum likelihood estimates of parameters of the general model and of relevant submodels are presented in Table IV. It is seen that the general path model considered in this study provides an adequate fit to the familial correlations observed for the standardized levels of both SBP and DBP. For standardized SBP, it is also seen that the subhypotheses of no cultural inheritance ($c = y = 0$, $i = v = 1$) and no extra sibling environmental effect ($b = 0$) are rejected at the 5% level. There is no evidence of any significant residual genetic inheritance of the standardized SBP, although under the general model the estimated values of residual genetic heritability in children (h^2) is about 20%, and that in adults (h^2z^2) is about 10%. No evidence of any effect of assortative mating and cohabitation, or of specific maternal environmental effect could be detected. The findings for the standardized DBP are broadly similar. Our interpretation of these findings is that the effects genetic factors (if any) on blood pressure levels is mediated through anthropometric characters such as those relating to obesity (weight and skinfold thicknesses); there is no residual effect of genetic factors on adjusted blood pressure levels.

TABLE IV

Maximum likelihood estimates of parameters and other statistics pertaining to systolic (SBP) and diastolic (DBP) blood pressure levels

Hypothesis	Variable	$-2 \ln L$	χ^2	*df*	h^2	c^2	*y*	*z*	*u*	*b*	f_F	f_M	*i*	*v*
General	SBP	2868.93	4.59	6	.20	.34	1.22	.84	−.12	.74	.19	.28	.84	.59
	DBP	2763.10	2.09	6	.15	.33	1.42	1.54	.33	.68	0.0	.27	.82	.69
$y = z = 1$	SBP	2869.25	0.32	2	.18	.36	[1.0]	[1.0]	−.07	.74	.21	.29	.82	.70
	DBP	2770.89	7.79	2.	.20	.22	[1.0]	[1.0]	.18	.61	0.0	.14	1.0	.92
$u = 0$	SBP	2869.25	0.32	1	.20	.35	1.10	.85	[0]	.74	.18	.28	.83	.65
	DBP	2767.44	4.34	1	.12	.33	1.54	1.34	[0]	.69	.08	.24	.81	.63
$b = 0$	SBP	2879.31	10.38	1	.15	.40	.87	.81	−.37	[0]	.62	.76	.78	.60
	DBP	2772.68	9.58	1	.15	.38	1.22	.93	.33	[0]	0.0	.50	.75	.80
$f_F = f_M$	SBP	2869.15	0.22	1	.20	.34	1.20	.84	−.12	.74	.23	.23	.84	.60
	DBP	2763.95	0.85	1	.12	.33	1.47	1.50	.30	.71	.14	.14	.82	.66
$h = z = 0$	SBP	2872.57	3.64	2	[0]	.43	1.13	[0]	−.08	.74	.25	.34	.75	.65
	DBP	2766.57	3.47	2	[0]	.40	1.34	[0]	.29	.69	.09	.30	.74	.72
$c = y = 0$, $i = v = 1$	SBP	2965.19	96.26	4	.65	[0]	[0]	.59	.10	.68	.08	.10	[1.0]	[1.0]
	DBP	2881.42	118.32	4	.48	[0]	[0]	.73	.13	.61	0.0	.11	[1.0]	[1.0]

Values in [brackets] indicate fixed values of parameters.

Logistic Regressive Model

While estimating the relative contributions of genes (major genes and polygenes) and environmental factors for a qualitative dichotomous character is somewhat more difficult under the classical liability threshold model[8], a more recent model uses a logistic regression formulation[1].

Suppose Y denotes a dichotomous random variable (r.v.) which takes the values 1 or 0 depending upon whether or not an individual possesses the character under study (e.g., affected or unaffected in respect of a disease). Let p = Prob $(Y = 1)$. Suppose C is a r.v. (possibly vector-valued) correlated with Y. The logistic regressive model assumes that

$$\text{logit } (p) = \log_e[p/(1-p)] = \alpha + \beta C = \theta. \quad \ldots (12)$$

Hence

$$p = \exp(\alpha + \beta C)/\{1 + \exp\{\alpha + \beta C)\}. \quad \ldots (13)$$

If $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$ denotes the observations on n members of a family, and $\Phi = (\boldsymbol{C}_1, \boldsymbol{C}_2, \ldots, \boldsymbol{C}_n)$ denotes the associated covariate ($\boldsymbol{C}_i$ is the covariate vector for Y_i), then we can write :

$$\text{Prob } (\boldsymbol{Y} \mid \Phi) = \text{Prob } (Y_1 \mid \Phi).\ \text{Prob } (Y_2 \mid Y_1, \Phi) \ldots$$

$$\text{Prob } (Y_n \mid Y_1, Y_2, \ldots, Y_{n-1}, \Phi) \quad \ldots (14)$$

For simplicity, we assume

$$\text{Prob } (Y_i \mid Y_1, \ldots, Y_{i-1}, \Phi) = \text{Prob } (Y_i \mid Y_1, \ldots, Y_{i-1}, \boldsymbol{C}_i). \quad \ldots (15)$$

In this model, it is further assumed that

$$\text{Prob } (Y_i \mid Y_1, \ldots, Y_{i-1}, \boldsymbol{C}_i) = \text{Prob } (Y_i \mid Y_F, Y_M, \boldsymbol{C}_i), \quad \ldots (16)$$

where $(F, M) \in (1, 2, \ldots, i-1)$ are the subscripts corresponding to the father and mother, respectively, of individual i. Hence,

$$\text{logit } (p_i) = \theta_i = \alpha + \beta \boldsymbol{C}_i + \gamma_F Y_F + \gamma_M Y_M. \quad \ldots (17)$$

By convention, $Y_F\,(Y_M)$ is zero if information on the father (mother) of the ith individual is absent in the data set.

Now, if $\boldsymbol{g} = (g_1, g_2, \ldots, g_n)$ denotes the ousiotypes (which is a more general term to denote major locus genotypes and/or polygenotypes and/or types that are culturally transmitted) of the n members, then

$$\text{Prob}(\boldsymbol{g}, \boldsymbol{Y} \mid \Phi) = \text{Prob } (\mathbf{g}).\text{Prob}(\boldsymbol{Y} \mid \boldsymbol{g}, \Phi). \quad \ldots (18)$$

$$\text{Hence,} \quad \text{Prob } (\boldsymbol{Y} \mid \Phi) = \Sigma_{\mathcal{P}}\ \text{Prob } (\mathbf{g}).\text{Prob } (\boldsymbol{Y} \mid \boldsymbol{g}, \Phi), \quad \ldots (19)$$

where the summation is over the set $\mathcal{P}$ of all possible vectors $\mathbf{g}$; $\Sigma_{\mathcal{P}}$ Prob $(\mathbf{g}) = 1$.

If the ousiotype refers to genotypes at a major locus with alleles A and a, then for any individual i who is a founder (i.e., with no parental data) in the family,

$$\text{Prob}(g_i) = \pi^2 \text{ for } g_i = 1 = AA$$

$$2\pi(1-\pi) \quad \text{for } g_i = 2 = Aa$$

$$(1-\pi^2) \quad \text{for } g_i = 3 = aa$$

where π is the frequency of the allele A in the underlying population which is assumed to be in Hardy-Weinberg equilibrium. If the individual i is not a founder, then

$$\text{Prob } (g_i) = \text{Prob } (g_i \mid g_F, g_M),$$

where g_F and g_M are the genotypes of the father and the mother, respectively. These probabilities are obtained using Mendelian transmission laws. The likelihood of the observations on a family are then computed using the Elston-Stewart algorithm[7] or some related variants (see Bonney[1] for further details).

Multilocus Recessive Model

Just as environmental contributions to the determination of a primarily genetic character adds to the "complexity" of transmission of the character, involvement of multiple genetic loci also leads to a similar complexity. If the character is not expressed at birth but at a later age, which is usually variable from one predisposed individual to another, there is further addition to the complexity. This arises because an individual possessing the predisposing genotype may not have manifested the character at the particular age when data are collected. During the past few years, we have been involved in the study of a dermatological disorder called vitiligo, which has a variable age at onset. This disorder is characterized by pale white patches on the skin, which tend to be progressive over time, and may eventually cover the entire body. The aetiology of vitiligo is unknown; a favoured hypothesis is that it is due to an autoimmune process. Reports of several twin pairs being concordant for vitiligo, and of many families exhibiting remarkable familial aggregation, lends credence to a genetic hypothesis. However, the disorder is found not to segregate in a simple Mendelian fashion, and although several environmental factors have been implicated, there is no compelling evidence of an environmental causation. The prevalence of vitiligo is roughly 0.5%. The modal age at onset is between 30 and 35 years. Some years ago, we (Majumder *et al.*[13]) have proposed a genetical model of vitiligo, based on family data collected by us from Calcutta. We have recently cross-validated the model using a set of fresh family data collected from the U.S.A. (Nath *et al.*[16]). In this section, we shall provide an overview of the model and the statistical method used in the analysis of these family data.

The model postulates that several (say L) unlinked, autosomal, diallelic (A_i and a_i denoting the two alleles at the ith locus) are involved in the pathogenesis of the disorder. Individuals of the genotye $(a_1\, a_1\, a_2\, a_2 \ldots a_L\, a_L)$ are said to be susceptible; and individuals of the remaining $3^L - 1$ genotypes are non-susceptible and never manifest the disorder. If, for simplicity, we assume that the frequency of the allele, a_i in the population is q, for all $i = 1, 2, \ldots, L$, and if the disorder manifests itself at birth, then the prevalence of the disorder in the population will be

$$\delta = q^{2L}. \qquad \ldots (20)$$

Properties of this model at the population level are derived in Li[11] and Majumder and Nath[12]. For the purpose of analyzing data on families, it is necessary to derive the likelihood of observations [phenotypes and ages (ages at onset for affecteds and

current ages for unaffecteds)] of members of a nuclear family. Since the data in these studies were collected through the presence of an affected individual in the family (which is a cost-effective strategy for sampling families in respect of a rare disorder), appropriate correction for ascertainment-bias is necessary. However, for analyzing data collected by ascertaining a family through an affected parent (the other parent being unaffected) no ascertainment-bias correction is required. The conditional likelihood function for phenotypic observations given the parental mating type (affected × unaffected) can be written as :

$$L = \Sigma_{k=1}^{K} \mu_k \Pi_{g=1}^{G} \binom{n_g}{m_g} [\theta_k (1 - z_g)]^{m_g} [1 - \theta_k (1 - z_g)]^{n_g - m_g} \quad \ldots (21)$$

where, m_g and $n_g - m_g$ denote the numbers of affected and unaffected offspring belonging to age group g (= 1, 2, ..., G), K denotes the number of possible genotypic matings corresponding to an affected × unaffected mating, θ_k denotes the probability that an offspring is of the susceptible genotype given that the parental genotypic mating is of type k (= 1, 2, ..., K), z_g denotes the probability that an individual in age group g (= 1, 2, ..., G) is phenotypically unaffected given that (s)he is of the susceptible genotype.

The conditional likelihood function needs to be corrected for ascertainment-bias when the family is ascertained through the presence of an affected offspring. For data on a family so ascertained, the corrected likelihood function of observations on children conditional on both parents being unaffected can be written as

$$L = [\alpha_M \cdot \lambda(\boldsymbol{n}, \boldsymbol{m})]/\beta(\boldsymbol{n}, \boldsymbol{m}) \quad \ldots (22)$$

where $\lambda(\boldsymbol{n}, \boldsymbol{m})$ is of the same form as the right-hand side of eqn. (21), the variables in this term are defined in a similar manner as before, $\boldsymbol{n} = (n_1, n_2, \ldots, n_G)$, $\boldsymbol{m} = (m_1, m_2, \ldots, m_G)$, $\alpha_M = 1 - (1 - \pi)^M$, $M = \Sigma_{g=1}^{G} m_g$, π = probability of ascertainment, and $\beta(\boldsymbol{n}, \boldsymbol{m}) = \Sigma_{r=1}^{N} [\alpha_r \Sigma_{\mathcal{P}} \lambda(\boldsymbol{n}, \boldsymbol{l})]$, $\mathcal{P} = \{\boldsymbol{l} \mid l_g \leq n_g \text{ and } \Sigma_{g=1}^{G} l_g = r\}$. If $\pi \simeq 0$, then $\alpha_M \simeq M\pi$, which results in some simplification of $\beta(\boldsymbol{n}, \boldsymbol{m})$. For further details, see Nath *et al.*[16]

Since other possible types of families were not observed in our data set (Nath *et al.*[16]), we shall not discuss these types. Our data set comprised information on 86 affected × unaffected families each ascertained through the affected parent, and 61 unaffected × unaffected families each ascertained through an affected offspring. For brevity, we shall not provide details of the distribution of age at onset of vitiligo which was used for the analyses of these data. These details can be found in Nath *et al.*[16]. In Fig. 2, are presented a plot of the values of the log-likelihood function separately for the affected × unaffected families, unaffected × unaffected families and the pooled set of all families. These values were computed using eqns. (21) and (22) for different numbers of loci. It is seen from this figure that for the entire set of data while there is a remarkable increase in the value of the likelihood function from 1 to 2 loci and from 2 to 3 loci, the subsequent increase in likelihood is minimal with an increase in the number of loci. Thus, our analysis reveals that vitiligo is a recessive disorder that is controlled by genes at three unlinked autosomal diallelic loci.

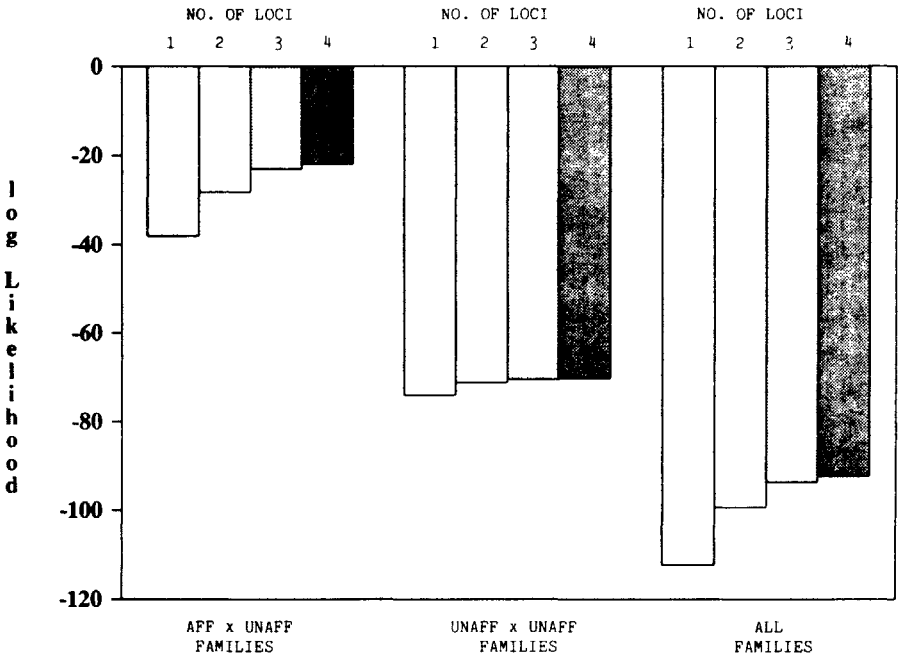

FIG.2. log-likelihood values of family data on vitiligo under the multiple recessive model for different numbers of loci.

ACKNOWLEDGEMENT

I would like to thank S. K. Bhattacharya, R. N. Das, S. K. Das, C. C. Li, B. N. Mukherjee, S. K. Nath, S. Nayak, J. J. Nordlund and D. C. Rao for collaboration and advise on the studies cited in this paper.

REFERENCES

1. G. B. Bonney, *Biometrics* **42** (1986), 611-25.
2. R. Chakraborty, *Canad. J. Cytology & Genetics* **19** (1977), 217-23.
3. R. Chakraborty, In : *Human Population Genetic : A Centennial Tribute to J. B. S. Haldane* (Ed. Partha P. Majumder), Plenum Press, New York, 1993.
4. R. Chakraborty and H. Danker-Hopfe, In : *Handbook of Statistics, Vol. 8: Statistical Methods in Biological and Medical Sciences* (Eds. C. R. Rao and R. Chakraborty), Elsevier Science Publishers B. V., Amsterdam. (1992), pp. 203-54.
5. R. Chakraborty and C. R. Rao, In : *Handbook of Statistics, Vol. 8 : Statistical Methods in Biological and Medical Sciences* (Eds. C. R. Rao and R. Chakraborty), Elsevier Science Publishers B. V., Amsterdam, 1992, 271-316.
6. C. C. Cockerham, *Evolution* **27** (1973), 679-700.
7. R. C. Elston and J. Stewart, *Human Heredity* **21** (1971), 523-42.
8. D. S. Falconer *Annals of Human Genetics* **29** (1965), 51-76.
9. J. Felsenstein, *Evolution* **39** (1985), 783-91.

10. C. C. Li, *Path Analysis : A Primer*, Boxwood Press, Pacific Grove, 1975.
11. C. C. Li, *Am. J. Human Genetics* **41** (1987), 517-23.
12. P. P. Majumder and S. K. Nath, *J. Genetics* **71** (1992), 89-103.
13. P. P. Majumder, S. K. Das and C. C. Li *Am. J. Human Genetics* **43** (1988), 119-25.
14. P. P. Majumder, R. N. Das, S. Nayak, S. K. Bhattacharya and B. N. Mukherjee, Genetic epidemiology of blood pressure in two Indian populations : Some lessons, *Human Biology* (in press).
15. N. E. Morton, D. C. Rao and J. M. Lalouel, *Methods in Genetic Epidemiology* S. Karger, New York, 1983.
16. S. K. Nath, P. P. Majumder and J. J. Nordlund, *Am. J. Human Genetics* **55** (1994), 981-90.
17. T. K. Nayak, Applications of entropy functions in measurement and analysis of diversity. Ph.D. Dissertation, Department of Mathematics and Statistics, University of Pittsburgh, PA, USA, 1983.
18. M. Nei, *Am. Naturalist* **106** (1972), 283-92.
19. M. Nei, *Genetics* **89** (1978), 583-90.
20. M. Nei, *Molecular Evolutionary Genetics*. Columbia University Press, New York, 1987.
21. M. Nei and R. K. Chesser, *Annals of Human Genetics* **47** (1983), 253-59.
22. M. Nei and A. K. Roychoudhury, *Genetics* **76** (1974), 379-90.
23. M. Nei, J. C. Stephens and N. Saitou, *Molecular Biology and Evolution* **2** (1985) 66-85.
24. J. Ott, *Analysis of Human Genetic Linkage*. John Hopkins University Press, Baltimore, 1985.
25. J. Ott, In: *Human Population Genetics : A Centennial tribute to J. B. S. Haldane* (Ed. Partha P. Majumder), Plenum Press, New York, 1993.
26. D. C. Rao, M. McGue, R. Wette and C. J. Glueck, In : *Human Population genetics* (Ed. A. Chakkravarti), Van Nostrand Reinhold, New York, 1984.
27. C. R. Rao, *Utilitas Mathematicas* **21** (1982), 273-82.
28. N. Saitou, In : *Handbook of Statistics, Vol.8 : Statistical Methods in Biological and Medical Sciences*, (Eds. C. R. Rao and R. Chakraborty). Elsevier Science Publishers B. V., Amsterdam, 1992, pp. 317-46.
29. L. D. Sanghvi, *Am. J. Physical Anthropology* **11** (1953), 385-404.
30. P. H. A. Sneath and R. R. Sokal, *Numerical Taxonomy*, W. H. Freeman, San Fransisco, 1973.
31. S. Wright, *Genetics* **28** (1943), 114-38.
32. M. Nei, *Proc. Natn. Acad. Sci., USA* **70** (1973), 3321-23.
33. D. R. Labarthe, *Annual Review of Public Health* **7** (1986), 193-215.