1 *Introduction*

The application of models based on surnames to the study of the genetic structure of the human population would seem to call for some justification. Any such application involves the assumption that the inheritance of surnames and biological inheritance are similar, or alternatively it must attempt to measure and allow for the differences between the inheritance of surnames and that of genetic traits.

One may begin to introduce the subject of surname models by an account of the scope of human biology, the place of human population structure in it, and the reason that models by analogy are needed. Human biology is concerned with the adaptive mechanism that makes human life possible. From one point of view this is controlled by those aspects of the genome that are shared by all humans and distinguish human beings from members of other animal species. Human life involves the response of human beings in various cultural and natural environments.

The other chief concern of human biology is human differences and the factors that account for them. Again these can be genetic at base, but they also involve interaction with the environment – which, for human beings, involves virtually the whole range of land habitats and is rendered much more varied still by the results of human activities and their variation from region to region.

The whole question of human biology, with complex diversity within the species overshadowed by the similarities among humans which distinguish *Homo sapiens* from other species, can be sketched synthetically but it cannot be studied as a whole. Instead, specific problems have to be isolated and attacked one at a time.

The landmark studies of human biology are of this kind. For instance, the classic report to a US Senate immigration committee by Franz Boas (dated 1910, but almost always cited as 1911) appeared at a time when the cephalic index was considered to be a hallmark of race and hence inherent and immutable. However, Boas showed that children reared to adulthood in a different country from that in which their parents were reared, and hence in a different environment, grow up to have, on average, a different cephalic index and a different stature. Boas' findings have since been

amply confirmed for Jews, Mexicans, Japanese, Chinese and other immigrant groups in the United States.

Another result which has proved to hold true in many subsequent investigations is Raymond Pearl's finding that tobacco smokers tend to live less long. Pearl had been associated with the eugenics movement and believed in a predominant role of genetics in human experience, including the determination of length of life. Probably it is no accident that someone with such a slant – that is, the belief that longevity is genetically determined – would successfully demonstrate the importance of an environmental factor. Those who need no convincing of the importance of environmental influences, on the other hand, may be expected to impose rigorous controls on them and produce good studies of the human genetics of quantitative traits.

All human biological attributes have both genetic and non-genetic aspects, of course, but that is not to say that specific observable differences are always the result of combined influences. In studying surname models, for instance, one is concerned purely with the genetic. Surnames do have non-genetic factors associated with their origin and mutation. These are peculiar to surnames, however, and are not like the non-genetic influences in biological mutation. That is, the origins of surnames are to be found in linguistic phenomena similar to those of other kinds of naming such as place naming. The rules for changing personal names in adoption, on marriage, or for other reasons are purely cultural rules. They must be understood and allowed for, not to include them in the models, but to minimise their influence on the models.

Illegitimacy has sometimes been cited as a condition in which the child has a different surname from its father and hence in which the assumptions in surname models of genetic transmission would be violated. In many countries, including England, however, the illegitimate child of an unmarried mother usually takes the mother's surname. If the purpose of modelling is to use the line of descent marked by surname as representative of all lines of descent (as is the case in the use of marriages between persons with the same surname for estimates of inbreeding) then lines passing through mothers are as valid as those passing through fathers, and children taking their mothers' surnames pose no problems. Illegitimate children of married mothers who take the surnames of their mothers' husbands do produce errors in the surname models, but the same instances usually cause the same problem in direct studies of the pedigrees. In recent years people in Western society have become more open in discussion of these issues of paternity and, by intensive interviewing, it might now be possible to estimate how many cases of 'wrong' surnames

are being included in a survey. In this way one could also recalculate coefficients derived from the survey to see how much difference these false classifications make. Until now, however, users of surname models have merely hoped and believed that, on average, the mothers' husbands and biological fathers are similar enough in relevant characteristics indexed by surnames (such as ethnic and geographic origins and consanguinity with the mother) for a few instances of mistaken paternity to have no significance.

One peculiarity of surnames is that there tend to be sharp discontinuities in the occurrence of particular surnames at national and other linguistic boundaries. These discontinuities may be sharper than are found in gene frequency distributions because, in the south and west parts of the European continent, present distributions of surname frequencies mark the results of migrations of approximately the last 40 generations whereas gene frequency distributions are the result of the ebb and flow of migrations over a very much longer, although indefinite, past. Thus the greater geographic variability of surnames than of human genes over the whole continent is not entirely – and perhaps not even importantly – due to the fact that pronunciation (and hence the way surnames are spelt) may be modified by local usage whereas genes remain in unaltered form among migrants and their descendants.

Surname models are important, then, for the very fact that they isolate genetic aspects and deal with them separately. The more one believes genetic–environmental interaction to be important in human biology, the more reason there is to start one's analysis in situations in which one or the other factor is minimized. One would stand to be criticized for this only if one considered human population genetics to be the whole of the science of human biology.

However, this is one aspect of surname models that has been criticized: what they do not do. Even within the genetic sphere, surname studies have not been (and cannot be) directly applied to the evaluation of natural selection. Natural selection is a mechanism by which genetic–environmental interactions influence genetic consequences, and the purely genetic approach of surname models is bound to miss this subject just as it misses questions of direct environmental impacts on human biology of the kinds posed by Boas and Pearl. Weiss and Chakraborty (1982) said: 'the selection–drift controversy is the central problem in evolutionary population genetics today. This is largely because there is no adequate selection model to explain the maintenance of the vast amount of polymorphism that has been found.' There are so many different specific distribution patterns of known human polymorphisms that no

single or few modes of selection would account for them all and some investigators attribute them to stochastic (chance) factors. Others think that random or stochastic merely means so far unexplained, but that all the patterns have their *raison d'être*. Weiss and Chakraborty, who seem to be speaking from the first of these positions, complain that the physical anthropologists who have studied human inbreeding by means of the prevalence of marriages between persons of the same surname, have pursued their studies with little sense of this problem. Indeed, the whole of the anthropological effort to study the breeding structure of human populations and of the species as a whole could be characterized as a descriptive enterprise.

This misses a main point of such studies, however, for there is an implicit programme in the study of surname genetics. Precisely because surnames cannot be subject to the forces of natural selection and must be considered neutral to selection by disease or climate, a difference (or lack of it) in geographic patterns of the tens of thousands of surname 'alleles' compared with those of biological alleles would be evidence bearing on what Weiss and Chakraborty call the central problem in evolutionary population genetics today: the selection–drift controversy. Unfortunately one cannot conclude that all differences between findings from surname distributions and those from studies of polymorphic genes are due to the action of selection in the latter case but absent in the former. It is also necessary to allow for the indefinite time span of biological genes and the rather limited history of surnames. This difference in itself permits approach through surname analysis to another important problem: the tempo of changes. Such estimates of the rate of change may be derived from surnames of several centuries' antiquity but be difficult from pedigrees of only three or four generations' depth and impossible in the equilibrium state reached by genes after an indefinite span of thousands of generations.

There are still other complications with surname analyses: the high frequency of multiple independent origins and mutations. Such shortcomings are not unique to this type of study; they are to some extent present in the alternative methods of approach. Even if the descriptive and historical information gained by surname analysis is not valued for itself, the light shed on systematic versus random selection warrants the addition of surname analysis (a relatively small additional effort) to other methods of study in human population genetics.

Whatever the merits of this line of thinking, the mere description of present and past population structure has value because of the political misuse to which faulty concepts can be put. History is fraught with

disasters caused by the misconception that the human population is formed of pure races and recent mixtures between them. The alternative conception, promoted by some extreme environmentalists, of a formless mish-mash, may be damaging too, because its rejection by the populace on the basis of their own experience may add fuel to, rather than dampen, the appeal of racialist views.

These various reasons may help justify the pursuit of surname models of population structure. Surname studies are by no means the end-all of human biology, which must involve the understanding of both genetic and environmental components of human variation and the interactions between them. However, surname genetics has a certain fascination which it is hoped will be conveyed by its explicit and implicit applications to questions that go beyond the origin, spread and extinction of surnames themselves. It is that fascination and the ready availability of research material that has led to the development of the body of knowledge on the subject.

Surnames are not distributed homogeneously in different places and among different social groups. The general purpose of surname studies in human biology is to measure the different probabilities of finding the same surnames in different times, places, groups and, especially, in marital partners. These probabilities can be compared with gene frequency distributions and assortative mating for genes of polymorphic systems. Similarities will allow one to use surnames to model the genes; differences may help in understanding the processes of differentiation of gene frequencies.

2 *History of surname studies in human biology*

Yasuda and Morton (1967) traced the history of the use of surname models for the study of human inbreeding to George Darwin's (1875) article in the *Journal of the Statistical Society*. Darwin's father, the famous naturalist Charles Darwin, and his mother, a member of the Wedgwood family of china pottery fame, were first cousins. Darwin was interested in the possible deleterious effects of consanguinity of parents and he wanted to know the frequency of cousin marriages in England. He therefore sought data on cousin marriages and on marriages between persons of the same surname in various sources such as *Burke's Peerage* and the Pall Mall social register. He then followed an ingenious line of thinking to estimate the proportion of marriages between first cousins. He reasoned that marriages to a person of the same surname who was not a first cousin would be proportional to the frequency of the surname in the population. This would be frequent only for common surnames. The Registrar General (1853) had published the frequency of the 50 most common surnames in the marriage registers and from the sum of the squares of these frequencies (0.0009207) Darwin estimated that marriages between unrelated persons of the same surname would be not much different from one per thousand. The excess over this of marriages of persons of the same surname was ascribed to cousin marriages and this was divided by the fraction of cousin marriages that were same-name marriages to give the number of cousin marriages in the population. Darwin concluded that the rate of first cousin marriages was about $4\frac{1}{2}$% among the aristocracy, $3\frac{1}{2}$% among the middle classes and landed gentry, $2\frac{1}{4}$% in the general population of rural districts and 2% in the cities.

Thirty-three years later, Arner (1908) published a similar study of cousin marriages in eighteenth-century New York and nineteenth-century Ohio. He first examined 10198 marriage licences in New York dated before 1784. The 50 commonest surnames gave an expectation of same-name marriages at random of 0.000757, but 211 of the marriages (0.0207) were actually between partners of the same surname. Using Darwin's ratio of same-name to all first cousin marriages would have led to an estimate of 5.9% cousin marriages in colonial New York. Arner

considered this estimate to be too high. He found genealogies in which there were 242 same-name marriages of which 70 were between first cousins. He believed that this rate might be biased by pedigrees in which only males were traced so he eliminated the all-male pedigrees and found that 24 of the remaining 62 same-name couples were first cousins. After rounding the ratios, Arner calculated that 2.76% of the marriages in colonial New York were between first cousins. In Ashtabula County, Ohio, he found records of 13309 marriages between 1811 and 1886 of which 112 were between persons of the same surname. He concluded from this, by Darwin's method, that 1.12% of these marriages were between first cousins.

Darwin's and Arner's studies were not immediately followed by others and they were apparently unknown to those who first revived study of the subject. In 1964 or so James F. Crow was invited to write an article on human inbreeding to accompany publication of one by Gordon Allen on random and non-random inbreeding. Crow recalled that in a lecture in the 1940s H. J. Muller (who received the Nobel Prize for his discovery that X-irradiation stimulates genetic mutations) had suggested that surnames could be used in genetic models of inbreeding. Crow (1983) says that he then had the thought for which his work on surnames is usually cited – that in most relationships the inverse of the likelihood of having the same surname times the degree of relationship is a constant number. Offspring of a brother and sister have an inbreeding coefficient of $\frac{1}{4}$ and always bear the same surname. Offspring of aunts and uncles with their nephews and nieces have an inbreeding coefficient of $\frac{1}{8}$ and bear the same surname in approximately one half of the instances, thus also yielding, on average, $\frac{1}{4}$ (that is $\frac{1}{8} \times \frac{2}{1}$) to the inbreeding coefficient of the population. Offspring of first cousin marriages have an inbreeding coefficient of $\frac{1}{16}$ and one type of cousin in four bears the same surname (fathers' brothers offspring, but not the offspring of fathers' sisters, mothers' brothers or mothers' sisters); thus the contribution to the inbreeding coefficient of the population represented by each married pair of same-surname first cousins is $\frac{1}{16} \times \frac{4}{1}$, which is also $\frac{1}{4}$. Crow took his idea and set forth this reasoning to a colleague, Charles Cotterman, who then worked on the problem of the possible relationships among four people and concluded that there are 17 such relationships and that 13 of them fail to agree with the 1 : 4 ratio. That is because across generations one's mother has a different (premarital) surname from oneself. Because of this theoretical difficulty Cotterman did not join Crow in writing the paper. Crow saw, however, that most consanguineous marriages are not across generations and that the four instances of the 1:4 ratio on Cotterman's list encompass most of the

instances in human marriages and hence in human mating. Crow also wrote to Muller to tell him about this development of his idea, but to Crow's astonishment Muller had completely forgotten and said he had never heard of such a thing.

Crow was interested in two prospects for the use of surname models: first, inclusion of remote common ancestry that would be revealed by a surname in common but missed in pedigrees from interviews; and second, separating the non-random component of inbreeding from the random one. The random component is the amount of inbreeding due to the limited size of the population breeding as though all possible matings had been equally probable. The non-random component is the extent to which this is increased (or decreased) by selective mating in a single generation. In order to work through these ideas on suitable data, Crow sought the cooperation of Arthur Mange, who had assembled records of the marriages of the Hutterites, a religious isolate living in the western United States and Canada. The Hutterites had collected much information about themselves and were willing to cooperate with Mange and other geneticists. Cotterman contributed further advice and Crow and Mange (1965) wrote about them in the most influential article ever to appear on the subject of surname models of human inbreeding. In it they estimated the inbreeding coefficient (in a population, the average proportion of genes at paired loci that are identical by descent from the same ancestors through both parents). Crow and Mange took the rate of inbreeding to be one quarter the frequency of marriages between persons of the same surname, which they called 'marital isonymy'. Furthermore, they partitioned the inbreeding coefficient into a random and a non-random component by a method they developed which adapts the approach of Wright (1922).

The idea of using surnames to study inbreeding was not new with Crow and Mange's publication. Although Crow and Mange were unaware of it, Yasuda and Morton (1967) knew of the studies by Darwin (1875) and Arner (1908) and that an American geneticist, Shaw (1960), had also pointed out that regular use of two surnames per person in Spanish-speaking countries provides an opportunity for effectively applying an index of consanguinity. In the Spanish system the given name or names is followed by the father's surname and then the mother's father's surname. Thus the last name is dropped each generation and replaced by a new name – that of the mother's father. Since married women usually retain their maiden names, this identifies an individual with both parents' families orientation (i.e. the families into which their parents were born and in which they grew up).

In the meantime (according to Yasuda, 1983), and without knowledge of the inbreeding coefficient developed by Wright (1922), Kamizaki (1954) had calculated the expected frequency of isonymy (i.e. having a surname in common) among various degrees of relatives (as Cotterman and Crow were to do later) and anticipated results of Crow and Mange. Kamizaki also cited earlier Japanese work in which the proportion of consanguineous marriages was estimated from isonymy. He reported too the same proportions of isonymy in consanguineous marriages as later derived by Crow ($\frac{1}{4}$ in first cousins, $\frac{1}{8}$ in first cousins once removed, $\frac{1}{16}$ in full second cousins, etc.) and derived general formulations for estimating the degree of relationship due to more remote consanguinity. Yasuda points out that change of a man's surname by adoption of his wife's, a frequent occurrence in Japan, will not ordinarily change the average frequency of isonymy. That is, it will shift the tested relationships from all-male lines to lines with female links, but will retain the same level of probability. This conclusion does not apply to the usual type of adoption in Western society, but does apply to cases of illegitimacy where the mother's surname is used for the child.

One other point should be made about work with Japanese surnames, however. Prior to the Meiji restoration, surnames were not allowed to be used except by the governing classes, and did not become mandatory until 1875. So it is only slightly over a century since surnames were arbitrarily assigned to almost all founders of the present surname lines. Thus, in Japan, inbreeding calculated from isonymy is for a period no more than about five generations – a length of time that can be encompassed by careful interviews and searches of family and other records. Since pedigrees include all consanguineous unions, but isonymy counts only a fraction and estimates the rest, over this span of time pedigrees provide the better way of determining the extent of inbreeding. On the other hand, the very fact that the time span of the use of surnames is about the same as that which can be covered in pedigree studies greatly enhances the value of a direct comparison of results by the two methods for evaluating the applicability of isonymy levels to estimating inbreeding.

Other studies in which several methods have been applied (among them studies in Switzerland and elsewhere by Morton and associates and by Ellis and associates, in the Pyrenees by Bourgoin and Vu Tien Khang and in the Orkney Islands by Roberts and Roberts) show rather poor correspondence of estimates of inbreeding from surnames and from pedigrees. In the West, unlike Japan, surnames have a considerable antiquity and the higher estimates of inbreeding from surnames than from pedigrees in some of these cases are at least partly explained by the

inclusion of remote inbreeding by the surname but not by the pedigree method.

Communities where there are few surnames some of which occur with high frequency must always have been known to be inbred. In 1957–8, during a comparative study of a number of communities on the north coast of Peru, I was aware of this and collected lists of surnames from various sources (interviews, birth registers, death registers and grave markers) and for various periods of time so that a scale of isolation or inbreeding could be devised and compared with rates of migration into the same communities. When it came to analysing the data, however, it was not clear how one could deal with the surname distribution data. Only much later, Wendy Fox, a mathematician, suggested that one could approximate the frequency distribution curve of surnames with a formulation that would permit comparison of the constants of curves fitted to different sets of data (Fox and Lasker, 1983). If the surnames of a population are listed by rank order of their frequency of occurrence, the log of number of surnames occurring x times against x tends to form a straight line with the log of number of occurrences. According to the reasoning of Zipf (1949) this would be so if the rate of growth or decline in frequency of a surname were independent of its frequency. The formula gives a good fit to some data from several adjacent areas in Berkshire and southern Oxfordshire (from marriages registered in Reading, Wokingham and Henley). Unfortunately the slopes of lines fitted in this way and of similar curves that can be fitted to surname frequency distributions are dependent on sample size. That had been apparent in the sets of data from Peruvian towns and villages. When the Crow and Mange (1965) article appeared a method was available for application to them that is unaffected by sample size (Lasker, 1968, 1969).

Recently, Zei *et al.* (1983*a*,*b*) have argued that the theoretical distribution of neutral alleles – the assumed model for surnames – is better matched by a logarithmic distribution originally introduced by R. A. Fisher to represent the variation in the abundance of species and applied to surname frequencies by Chakraborty *et al.* (1981). Zei *et al.* (1983*b*) compared the fit of these different formulations and in some examples found their method seems to describe the data more precisely than the method of Fox and Lasker. Furthermore, they showed how to take sample size into account and published a table for use in the necessary computations. Wijsman *et al.* (1984) have further extended these studies by developing a method for calculation of migration rates from the matrices of surname frequencies in the various places in an area at two or more periods of time.