J. Genet. Vol. 71, No. 3, December 1992, pp. 81–87. © Printed in India.

Haldane and the mutation rate*

J. H. EDWARDS
Genetics Laboratory, Biochemistry Department, University of Oxford, South Parks Road, Oxford OXL 3QU, UK

MS received 6 November 1992

An obvious implication of Mendel's atomism, that is, that the units of inheritance usually pass unchanged from flower to seed, or parent to child, and of Darwin's single evolutionary tree of life was that differences in these atomic units were both a cause of individual differences and the raw material of speciation. Haldane's greatest contribution was to consider the changes to these individual atoms in their nature, their number and their organization.

The simplest situation, that in which two organisms differ only in the nature of these units, or genes, their number and arrangement being the same, was the basis of Haldane's seminal papers in which he explored the consequences of a change or mutation in a single gene. The gene continued, as the rider in the postal system of ancient Rome who would stop at a *mutatio* to change horses. It was this apparently simple problem, a problem basic to the whole of evolution, that dominated Haldane's work in genetics in the forty years bounded by his first paper in the series on a mathematical theory of natural and artificial selection (Haldane 1924) and his "Defense of Beanbag Genetics" (Haldane 1964). Throughout he was interested both in natural events and in the consequences of artificial induction of mutations by the inevitable hazards of medical diagnosis and therapy, the necessary hazards of some industrial processes, and the unnecessary hazards of excessive medical activities, of casualness in commerce, and of both war and the testing of the weapons of war.

His first paper on a genetic subject, on colour variants in rats, referred to by Mitchison (1968), is not quoted in any bibliography known to me. His second, on the linkage of two colour variants in mice (Haldane *et al* 1915), was with his sister Naomi, later Naomi Mitchison, and Sprunt, published after Sprunt's death in the trenches; the first paragraph included a sentence ending "one of us is already dead". This was possibly the first mammalian linkage to be described although at that time the observations were thought to represent a change in number and termed reduplication. The nature of the more severe form of albinism, common to man and mouse, is now finally resolved at the DNA level while the nature of the second locus, pink-eyed dilute, has only just been resolved (Lyon *et al.* 1992) and its probable human homologue, tyrosine-positive albinism, was only published (Ramsay *et al.* 1992) a month before his centenary. It is not clear whether Spooner, the Warden at New College, at which Haldane and his father were Fellows, suffered from tyrosine-negative

* Based on a talk given at the Haldane Centenary Symposium held on 6 November 1992 at Ahmedabad as part of the 58th Annual Meeting of the Indian Academy of Sciences.

or tyrosine-positive albinism. The extreme whiteness of hair in his portrait favours the former but appearances can be deceptive. He always denied spoonerisms.

Haldane's fourth genetic paper (1919), published after returning to Oxford from New Delhi where he probably wrote it, was on the consequences of rearrangement, following the key papers by Sturtevant (1913) on how to order loci, by Robbins (1918) on allelic association in populations due to very close linkage, and by Fisher (1922) on how to optimize the position of ordered loci, the first application of maximizing likelihood by a method he later formalized as inversion of the information matrix.

In this paper Haldane related the functional or genetic length of a chromosome, determined by the cumulative probability of recombination between neighbouring units, to structural length on the assumption of two strands and a random distribution of crossovers, introduced the mapping function now known by his name, advanced and defined the term morgan, and consolidated the occasional use by Bateson and Punnett of Mendel's upper and lower case to define the origin, rather than the state, of an allele. He introduced an empirical modification of his basic equation

$$y = (1 - e^{-2x})/2$$

relating recombination fraction (y) to distance (x), but it was left to Kosambi (1944), an Indian mathematician whose width of interest included the death-rates of coins, to carry it further. Kosambi's function, which includes the intellectual *frisson* of homology with the addition rules for velocity in general relativity, is now commonly used although it fails to allow for the rarity of close, as opposed to distant, double recombinants.

The next substantial contribution to the mathematics of linkage, after extensive groundwork by Fisher, Yates, Mather and Bailey, was that of Morton in 1955. We desperately need another paper of substance either to consolidate the foundations of much recent work, or, if faulty, to rebuild the foundations or lighten the superstructure. As Fisher pointed out, linkage is so difficult a subject that we must constantly check the mathematics involved against reality. We now have a discrepancy of the order of two between the length of much of the human genome based on microscopy at meiosis and the length inferred from various methods of multilocus mapping and automated ordering.

Haldane (1956) was the first to emphasize the need to allow for a multiplicity of tests in assessing the significance of linkage, a need ignored in Botstein *et al.* (1980) in their paper spawning the use, once again, of the term the "New Genetics", an accolade initially awarded to the great work of Morgan and his collaborators and later, to that of Lysenko. This precaution is widely disregarded and has done much to burden the human gene map with dubious locations of the various non-Mendelian loci predisposing to psychosis. Haldane, using the useful term "false clue", pointed out that with 25 markers a linkage value over two standard deviations from the null value of 1/2 would be expected in $1 - 0{\cdot}97725^{25}$ or 47·2% of studies.

Haldane's first paper on "A Mathematical Theory of Natural and Artificial Selection" (1924), written from Cambridge where he was Reader in Biochemistry under Gowland Hopkins, started "A SATISFACTORY theory of natural selection must be quantitative" (his capitals), and set out the programme for his future papers. He stated:

> In any given case we must specify:
> (1) The mode of inheritance of the character considered.

(2) The system of breeding in the group of organisms studied.
(3) The intensity of selection.
(4) Its incidence (e.g. on both sexes or only one), and
(5) The rate at which the proportion of organisms showing the character increases or decreases.
It should be possible to obtain an equation connecting (3) and (5).

Since a mutation is, by definition, an event leading initially to an incidence of $1/n$ gametes, or $1/2n$ individuals if diploid, this specified the various problems he was later to solve, elucidate, and sometimes demonstrate to be insoluble by analysis.

Although the first examples only dealt with "the simplest possible case, a completely dominant Mendelian character", its development rapidly uncovered the limitations of the mathematical approaches available. He could not then, as we can now, delegate a million generations of simulation to a desk computer which could complete its work in a coffee-break. Ten iterations could be a morning's work, and about as much as anyone whose intellect was so powerful could be expected to tolerate. He was well aware of these difficulties before he started, and stated:

> Even so it will be found that in most cases we can only obtain rigorous solutions when selection is very rapid or very slow. At intermediate rates we should require to use functions of a hitherto unexplored type. Indeed the mathematical problems raised in the more complicated cases to be dealt with in subsequent papers seem to be as formidable as any in mathematical physics. The approximate solutions given in this paper are however of as great an order of accuracy as that of the data hitherto available.

This prediction has been validated. The diffusion equations of Kolmogorov have resolved some of the problems in the hands of Kimura and others, but can only satisfy those who are at ease with the use of a real number to summarize in one number the ratio of two integers of very different magnitudes and the use of the luxury of infinitesimals to navigate through a finite and often sparsely populated universe.

Throughout he emphasized the power and potential of artificial selection as a possible partial alternative to the tragedies natural selection imposes on our own species, but, unlike many contemporary advocates of artificial selection, he always combined brevity and clarity in word with rigour in mathematics. He suggested the use of linkage in predicting the presence of disorders not yet manifest in the first paper on linkage in man, mentioning Huntington's chorea (Haldane and Bell 1937).

Clarity was, by necessity, hardly a feature of some of his mathematics. This was in part due to his facility to expand where he could not integrate and his obsession of giving many terms in the many infinite series he partially enumerated. As with *Principia Mathematica,* whose authors are said to have read the proofs while no-one read the text, it is doubtful if anyone has read the full text down to the last subscript and the last term of every expansion in Haldane's many papers covering both genetics and statistics. *Principia Mathematica* has now been read by computers, and the very few errors or inconsistencies detected. Haldane's work awaits this perhaps superfluous validation, for while he aimed at precision he appears also to have used symbolism to convey both general trends and the impossibility of certainty.

I do not know if he would have appreciated the recent invasion of chaos as a theoretical justification for rough but robust approximations if supported by clarity in words, but he would have appreciated the power of such approaches to generate visual patterns which combine beauty and obscurity, especially as he claimed to lack any appreciation of music.

This use of the integral values and the finite difference calculus, while faithful to nature, was resistant to art, and the analytical solution of many of the simplest equations is usually difficult and often demonstrably impossible. Haldane resolved this by extensive trial on imaginary data as supports to his intuitive approximations and biological expectations, and his hopes that the leading elements in any series dominated the whole.

Mutation, or the change in state, and usually in function, of a gene might seem to lead to a simple billiard-ball type of model so effective in classical physics in which, because the model is simple to specify and reasonably faithful to reality, the consequences would be simple to deduce. But this is not so: a mutation starts as a unit, an atom amongst thousands, while the universe of classical physics usually starts with all the units being indefinitely numerous.

We are at once up against the central problem of applied mathematics. We are dealing with unit, or countable, phenomena. Although the mathematics of units, or integers, is simple at the level of addition, subtraction and multiplication, and the first part to be taught to children, further elaborations rapidly lead to difficulties and problems which are simple to specify, such as the next prime number, are often impossible to solve except by enumeration. In practice in almost all mathematical operations outside the simplest and cheapest financial transactions we usually work with real numbers, that is, numbers which, aided by decimals, can take any intermediate value between neighbouring integers, and these numbers are used as approximations to the naturally exact integers.

Haldane starts his study of selection using ratios, rather than proportions, with a population of gametes of two phenotypes A and B in the ratio of p:1. These then undergo fertile unions in the ratio of p: $(1-k)$ by which he defined k, his coefficient of selection, a number which can be positive or negative. His use defines it as a measure of fitness. This may seem at variance with the majority of mutations but Haldane was discussing advantageous mutations which exert their relative advantages by invasions, conquests, tolerance to disease or novel foods. He later justified this mathematical fiction of a gametic phenotype in his "defense of beanbag genetics", perhaps his finest conjunction of words and numbers.

Haldane next considers, a mere page later, "family selection", stating, "There is however another type of selection which so far as I know has not been considered in any detail by former authors, but which must have been of considerable importance in evolution." By family selection he defines selection within a niche whose resources are fully utilized and which is largely populated by close relatives, pointing out both the very high selection likely and its limited effect due to the genetic similarity of winners and losers.

The general formula, in Haldane's terminology, is

$$u_{n+1} = u_n + f(u_n),$$

where u is some number or proportion relating to the number of generations defined by a subscript and $f()$ is some function defining how each generation changes. He introduced this in its simplest form as

$$u_{n+1} = u_n/(1-k), \text{ and}$$
$$u_n = (1-k)^{-n} u_0.$$

Haldane resolves the difficulty in starting with the very discrepant proportion of

$1/n$ for the initial generation by the ingenious approach, also used by Newton in his analysis of scales and harmonics, of starting in the middle. The numbers of the mutant and "wild type" alleles are then equal, and, as there is symmetry to a close approximation if codominant, and recessive and dominant manifestation show a mutual symmetry, this treatment leads to a sigmoid curve approximating the cumulative normal distribution and clearly emphasizing that, in time, the main delays in establishing a new allele are shortly after its birth and shortly before the death of its partner, its distant parent. The time-scale is therefore dominated by n, the population size, or effective population size, the two great unknowns in any species in the wild, and not easily interpreted in the most documented human populations.

In four pages Haldane has clarified the key issues in "natural and artificial" selection with a few equations of rapidly deepening complexity although only approximate. The remainder of this series discusses the differences, which are mainly of scale rather than nature, related to various forms of inequality of representation in diploids, that is, of deviations from the bean-bag model of the sickly gamete.

These models include direct gametic selection both between gametes from diverse sources, as in wind-blown pollen, and from the same source, as is usual in internal fertilization, and cover selection in such species as bees which are dependent on the efforts of sterile workers and idle males, introducing what he later termed "altruism", the origin of what is now often termed "inclusive fitness".

In a later paper (Haldane 1927a) he extends the argument from the simpler case of assuming discrete generations, as in many plants and insects, to that of populations with overlapping generations, assumed to reproduce at an even rate, a close approximation in our species. This considerably, and perhaps needlessly, complicates the simple formulation but allows an easier transition to the use of real numbers and their infinitesimal aids. The next paper (Haldane 1927b) starts "New factors arise in a species by the process of mutation. The frequency of mutation is usually small but it seems probable that it can sometimes be increased by changes in the environment", a concept later to dominate his work on the hazards of radiation, a field in which both his predictions and the procedures he recommended for their refinement provide the basis for the present arbitrary but necessary levels of permitted exposure.

He follows Fisher in using the simplification that the unknown is randomly defined so that family size is Poissonian, but extends this to the effects of selective advantage, on the grounds that "if a large number of offspring is possible, as in most organisms, the series approximates to a Poisson series provided adult organisms are counted". This would seem to involve a somewhat unrealistic oversimplification since the mean number of adults in the next generation must average about two if population number is to be stable, as is assumed. At least this simplification leads to the elegant approximation

$$k = y/2 + y^2/3 + y^3/4 + \ldots,$$

where $(1 - y)$ is the chance of extinction and k the selective advantage. This result, that the chance of fixation of a new dominant is about twice its selective advantage, leads to his next result that if a mutation with such an advantage appears more than $\log_e(2)/(2k)$ times, or about $1/(3k)$, then fixation is to be expected. That is, a mutation with a selective advantage of 1/30 which, if it occurred a hundred times, would have an even chance of becoming established.

This is a very powerful result, and even if the approximations necessary for its

derivation lead to an error of a factor of two, which seems likely, or even a factor of ten, then this is of little import. Ten million and a hundred million years, the range of most mammalian speciations, are both adequate as a background on which to imagine the action of evolution. At least it makes clear that the inevitability of gradualness can only be fuelled by recurrent episodes of change, and, with mutation rates, which he quotes in the same paper as of the order of one in a million, an estimate which still stands, at any rate as a geometric mean, there is an immediate "world enough and time" problem. A million organisms in a million years would only have two million mutations and if as many as one in twenty were advantageous at a 1 in 30 level there would only be time to establish a thousand new advantageous alleles. However, their cumulative increase in selective advantage, should they compete with their ancestors of a mere million years ago, would be decisive. Later he was to introduce the darwin as a unit of evolutionary measure, defining it as a change in proportion of a millionth a year.

This formulation implies no probability of fixation of a neutral gene and breaks down with disadvantageous genes when k is positive. It is unlikely that at this time he envisaged the gene number as outside the thousand to ten-thousand range. He was to address this six years later in "The part played by recurrent mutation in evolution" (Haldane 1933). With characteristic lucidity he wrote, "Lethal genes, like parasites or predators, are part of the environment of other genes, and the necessary occasional stowaway in the establishment of any genome."

In 1957, when he presented one of the two papers I was privileged to hear at a meeting of the Genetical Society in Sheffield, he showed, by a very simple if approximate argument, which largely followed from his 1924 paper, that the cost of the replacement of an allele by its partner—a sort of genetical parricide—involved death by selection of numbers exceeding the standing population severalfold if individuals were numerous and matings random. Again an error of even a hundred-fold from such approximations does little to destroy the image of extreme gradualness of evolution unless aided by various catastrophic events in the cell, such as duplication and tetraploidy, which he had already considered (Haldane 1927c; see also Searle 1964), or through the creation of small isolates by changes in climate or geography.

References

Botstein D., White R. L., Skolnick M. and Davis R. W. 1980 Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* 32: 314–331

Fisher R. A. 1922 The systematic location of genes by means of crossover observations. *Am. Nat.* 56: 406–411

Haldane J. B. S. 1919 The combination of linkage values and the calculation of distances between the loci of linked factors. *J. Genet.* 8: 299–309

Haldane J. B. S. 1924 A mathematical theory of natural and artificial selection. Part I. *Trans. Cambridge Philos. Soc.* 23: 19–41

Haldane J. B. S. 1927a A mathematical theory of natural and artificial selection. Part IV. *Proc. Cambridge Philos. Soc.* 23: 607–615

Haldane J. B. S. 1927b A mathematical theory of natural and artificial selection. Part V. Selection and mutation. *Proc. Cambridge Philos. Soc.* 23: 838–844

Haldane J. B. S. 1927c The comparative genetics of colour in rodents and carnivora. *Biol. Rev.* 2: 199–212

Haldane J. B. S. 1933 The part played by recurrent mutation in evolution. *Am. Nat.* 67: 5–19

Haldane J. B. S. 1956 The detection of autosomal lethals in mice induced by mutagenic agents. *J. Genet.* 54: 327–342

Haldane J. B. S. 1957 The cost of natural selection. *J. Genet.* 55: 511–524

Haldane J. B. S. 1964 A defense of beanbag genetics. *Persp. Biol Med.* 7: 343–359

Haldane J. B. S and Bell J. 1937 The linkage between the genes for colour-blindness and haemophilia in man. *Proc. R. Soc. London* B123: 119–150

Haldane J. B. S., Sprunt A. D. and Haldane N. M. 1915 Reduplication in mice. *J. Genet.* 5: 133–135

Kosambi D. D. 1944 The estimation of recombination values. *Ann. Eugen.* 12: 172–175

Lyon M. F., King T. R., Gardo V., Gardner G. M., Yoshomicih N., Eicher E. M. and Brilliant M. H. 1992 Genetic and molecular analysis of recessive alleles at the pink-eyed dilution locus of the mouse. *Proc. Natl. Acad. Sci. USA* 89: 6968–6972

Mitchison N. 1968 Beginnings. In *Haldane and modern biology* (ed.) K. R. Dronamraju (Baltimore: Johns Hopkins Press) p. 303

Morton N. E. 1955 Sequential tests for the detection of linkage. *Am. J. Hum. Genet.* 7: 277–318

Ramsay M., Colman M.-A., Stevens G., Zwane E., Kromberg J., Farral M. and Jenkins T. 1992 The tyrosine-positive oculocutaneous albinism locus maps to chromosome 15q11.2-212. *Am. J. Hum. Genet.* 51: 879–884

Robbins R. B. 1918 Some applications of mathematics to breeding problems III. *Genetics* 3: 375–389

Searle A. G. 1968 Coat color genetics and problems of homology. In *Haldane and modern biology* (ed.) K. R. Dronamraju (Baltimore: Johns Hopkins Press) pp. 27–41

Sturtevant A. H. 1913 The linear arrangement of six sex-linked factors in drosophila, as shown by their mode of association. *J. Exp. Zool.* 14: 43–69