1-1-1996

Monte Carlo experiments and the defense of diffusion models in molecular population genetics

Michael Dietrich
Dartmouth College

Biology and Philosophy **11:** 339–356, 1996.

Monte Carlo Experiments and the Defense of Diffusion Models in Molecular Population Genetics

MICHAEL R. DIETRICH
History and Philosophy of Science Program
University of California, Davis
Davis, CA 95616
U.S.A.

Abstract. In the 1960s molecular population geneticists used Monte Carlo experiments to evaluate particular diffusion equation models. In this paper I examine the nature of this comparative evaluation and argue for three claims: first, Monte Carlo experiments are genuine experiments; second, Monte Carlo experiments can provide an important means for evaluating the adequacy of highly idealized theoretical models; and, third, the evaluation of the computational adequacy of a diffusion model with Monte Carlo experiments is significantly different from the evaluation of the empirical adequacy of the same diffusion model.

Key words: Experimentation, idealization, Monte Carlo methods, neutral theory of molecular evolution.

Introduction

J.B.S. Haldane once wrote that "Men have fallen in love with statues and pictures. I find it far easier to imagine a man falling in love with a differential equation, and I am inclined to think that some mathematicians have done so" (Haldane 1933, p. 30). After having spent months working through Motoo Kimura's papers, I am inclined to think that some population geneticists have done so as well. In Kimura's case, a particular class of differential equations called diffusion equations have long held his interest (Kimura 1985, p. 20). Since R.A. Fisher first introduced them into population genetics in 1922, diffusion equations have proven to be extremely useful and provide much of the foundation for stochastic population genetics (Fisher 1922). If you were a population geneticist with an interest in random processes, as Kimura was, diffusion equations were good things with which to be smitten.[1]

Unfortunately, diffusion equations have their shortcomings and carefully determining their applicability is an important part of their use (Kimura 1964, p. 181; Gillespie 1989, p. 57). In the 1960s, Kimura, Warren Ewens, and others began to put stochastic population genetics on a firmer footing by devising experiments to evaluate specific diffusion equation models (diffusion models). Kimura's experiments, however, were not done in a lab with beakers,

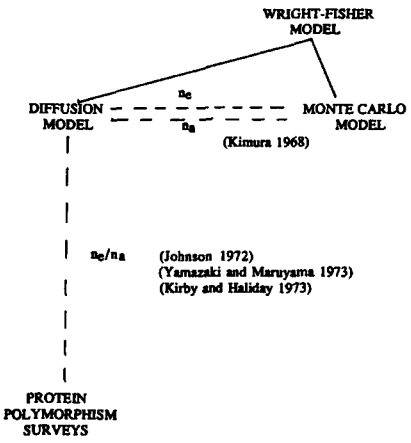

Fig. 1. Relations Among Versions of the Infinitely Many Alleles Models and Laboratory Models. Solid lines depict representation relations. Dashed lines depict comparative relations between models in terms of the values produced by each for the variables appearing next to the lines. Citations for these comparative evaluations are given in parentheses.

testubes, and hordes of *Drosophila*. They were done with punch cards in a room with an IBM 7090 computer (Kimura 1968b, p. 253).

The experiments Kimura, Ewens, and others conducted were called Monte Carlo experiments. They were used in the 1960s to evaluate specific versions of the infinitely many alleles model.[2] In this paper, I would like to demonstrate three things with regard to this kind of comparative evaluation in science: first, Monte Carlo experiments are genuine experiments, second, Monte Carlo experiments can provide an important means for evaluating the computational adequacy of highly idealized theoretical models, and lastly, this type of evaluation of theoretical models with Monte Carlo experiments is significantly different from the later evaluation of the empirical adequacy of that same theoretical model. The details of these Monte Carlo experiments and infinitely many alleles models will be discussed below. For now I want to present a general sketch of the comparison involving Monte Carlo experiments and diffusion models in order to better frame the issues which will be the concern of this paper.

The basic situation can be represented in terms of three different infinitely many alleles models. The idea behind any infinitely many alleles model is usually characterized as the supposition that there are a sufficiently large number of possible allelic states such that every mutation event creates a previously non-existent allele. The first infinitely many alleles model uses a Markov chain process to represent changes in allele frequency over time. In a Markov chain the probability that the frequency of an allele will have a

certain value is dependent on the values for allele frequency of the immediately preceeding time but not for any other times in the past. This model is probabilistic but it is discrete. This type of model is often referred to as a Wright-Fisher model after Sewall Wright and R.A. Fisher who pioneered the use of this kind of Markov chain model though not with the infinitely many alleles assumption. A second version of the infinitely many alleles model is the model inherent in the Monte Carlo experiment. This model is meant to represent the Wright-Fisher model and shares many of its features, but has other features that allow it to be instantiated in a computer which can then compute values for certain variables given a supply of random numbers. The third model is the diffusion model. This is also meant to represent the Wright-Fisher model, but does so in a different way from the Monte Carlo model. The diffusion model is a mathematical model using differential equations know as diffusion equations or Kolmogorov equations. The Wright-Fisher model and the Monte Carlo model both represent time and allele frequency as discrete variables, while the diffusion model represents time and allele frequency as continuous variables. The diffusion model is, thus, an approximation of the Wright-Fisher model. Because the Monte Carlo model retains allele frequency and time as discrete variables, it is thought to be a better representation of the Wright-Fisher model than the diffusion model (see Figure 1).

When Monte Carlo experiments are used to evaluate diffusion models, values predicted by the diffusion model are compared with values generated by the Monte Carlo experiment. What is at stake is the ability of the diffusion model to match the results of the Monte Carlo experiment. Too great a difference between the results of the diffusion model and the results of the Monte Carlo experiment and the diffusion model could be judged to be computationally inadequate.

The need to evaluate diffusion models

In the 1960s as more mathematicians and mathematically oriented biologists began working on diffusion models, greater efforts began to be made to check the adequacy of these models for different situations (Watterson 1962; Ewens 1963, 1964a,c). As computers and Monte Carlo experiments became more widespread they too began to be used to check diffusion models (Ewens 1966). Kimura was aware of these efforts from the early 1960s and in 1964 argued that because diffusion models were approximations "based on rather intuitive arguments", it was an "important task for mathematicians" to "investigate the conditions under which such approximations may be valid" (Kimura 1964, p. 181).

To understand the need to evaluate diffusion models we have to take a closer look at how they were being used in population genetics. If you are concerned with how random drift will affect the frequency of some allele over time, then you will want to build a stochastic model where allele frequency at any given time is not represented by a single value but by a probability distribution.[3] The probability distribution captures the fact that because the process of change in allele frequency is probabilistic, there will be a number of possible outcomes at some later time each with a certain probability of occurrence. The diffusion equation provides a way of representing the changes in probability distributions such that, given the initial allele frequency, the probability that the allele frequency takes the value x at time t can be determined for any time t (Gale 1990, p. 347). Diffusion models are very powerful tools, but they make several approximating assumptions.

A defining feature of diffusion models is that they describe discrete processes as continuous processes, although all diffusion processes require that changes be Markovian (Gale 1990, p. 45). Where changes in allele frequency had been described as occurring in discrete steps, 0, 1/2N, 2/2N, ..., (2N-1)/2N, 1, where N is population size; diffusion equation models treat allele frequencies as varying continuously from 0 to 1 (Kimura 1955, p. 144). Similarly, where time had been represented in terms of discrete generations, diffusion models treat it as a continuous variable. The "intuitive argument" for these changes is that evolution is usually a very slow process and the populations involved are usually very large, so differences in time and allele frequency become very small and can be approximated by continuous variables. Treating changes in allele frequency as a continuous stochastic process means that as "the time interval becomes smaller so does the amount of change in gene frequency x during that interval" (Kimura 1964, pp. 181–182).

Adopting the diffusion equation approach and making these approximations involves trading off accuracy for tractability (Gillespie 1989, p. 57). Kimura and his colleagues were aware of this in the early 1960s and wanted to be able to assess the accuracy of diffusion models as a result. Diffusion equations were assumed to be widely applicable, as long as the mutation rate was not of a larger magnitude than 1/N, where N is population size (Karlin and McGregor 1964, p. 255; Ewens 1964b, p. 892), but the adequacy of diffusion models in these applications was not formally supported. The need to check these extremely powerful models motivated the practice of testing with Monte Carlo experiments.

For Motoo Kimura, however, the stakes were higher than just bolstering a favorite mathematical approach to modeling. Kimura's test of diffusion approximations in 1967 (published in 1968) coincides with the beginning of his advocacy of the neutral theory of molecular evolution. At a time when

selection was thought to be the only important factor in evolution, Kimura had used a familiar cost of selection argument from evolutionary genetics to argue that the rate of change in large biological molecules was too great to be sustained, if most of those changes were harmful. A large number of harmful changes would be too great a burden on the fitness of the population and would drive it toward extinction, unless most of the changes were neutral, i.e., neither highly beneficial nor highly detrimental (Kimura 1968a). A large number of neutral alleles was, according to Kimura, how populations avoided what would otherwise be an excessive genetic load.[4] The factor governing such a system of neutral alleles would be random drift. The idea that most changes detected at the molecular level are neutral or nearly neutral and so governed by random drift is now recognized as the core of the neutral theory of molecular evolution. With his advocacy of the neutral theory, Kimura effectively reintroduced random drift as a biological factor to be reckoned with at the molecular level. It is in this context that Kimura used Monte Carlo experiments to check the diffusion equations underlying an important part of the neutral theory, the infinitely many alleles model.

Kimura had formulated the basic infinitely many alleles model in 1964 with James Crow (Kimura and Crow 1964). Infinitely many alleles models are models of mutation that posit that there are a vast number of alleles that could be formed so each mutation is a mutation to a new allele. The model was based on Joshua Lederberg's suggestion to Crow in 1958 that molecular genetics would soon show that every mutant was distinguishable from every other mutant (Crow 1989, p. 631).

Kimura and Crow were interested in the situation where the number of new alleles introduced by mutation balances the number of old alleles lost by random drift. This equilibrium situation should produce fairly constant values for variables such as the effective number of alleles and the average number of alleles maintained in the population. In the diffusion equation version of the infinitely many alleles model, this situation could be approximated by a stationary distribution.[5] The major result of the basic Wright-Fisher version of the infinitely many alleles model for neutral alleles was the prediction that the effective number of neutral alleles, n_e, capable of being maintained in a population is

$$n_e = 4N_e u + 1, \tag{1}$$

where N_e is the effective population size and u is the mutation rate.[6] Although this result was not crucial to Kimura and Crow's conclusions in 1964, it became quite important after the introduction of the neutral theory in the late 1960s and, especially, after the issue of the importance of neutral alleles in nature became highly controversial.

When Kimura returned to infinitely many alleles models in 1967, his goal was to put it on a firmer footing by evaluating the adequacy of the diffusion equations used in its construction. Kimura rederived the results of the infinitely many alleles model with the diffusion equation approach, computed predicted values for the effective and average numbers of alleles, n_e and n_a, respectively, and then compared these predicted values with the values for n_e and n_a generated by Monte Carlo experiments. This same procedure had been followed by Warren and P.M. Ewens in 1966 and undoubtedly formed the foundation for Kimura's analysis (Ewens and Ewens 1966, cited in Kimura 1968b, p. 253). However, where the Ewenses only did one Monte Carlo experiment to test their diffusion model, Kimura did several.

Monte Carlo experiments were not the only way to evaluate the approximations in diffusion models. Warren Ewens had also pioneered a method of evaluating the accuracy of diffusion models using an analysis of the leading terms in a Taylor series expansion (Ewens 1964a, 1965). Ewens' analysis provided correction terms for measured differences between values from diffusion models and the true values. Kimura was aware of this alternative approach and chose to use Monte Carlo experiments to evaluate the accuracy of his diffusion models probably because of their ease of application to dioecious populations.[7]

Monte Carlo experiments

In the kinds of Monte Carlo experiments done by Kimura and the Ewenses, random numbers were used as the basis for decisions about what value a specific variable was going to take.[8] The basic technique can be illustrated by considering a simple coin toss. The variable of interest is how the coin lands – heads up or tails up. In a computerized Monte Carlo experiment, a random number generator and computer algorithm would be used to determine which value this variable takes – whether it is heads or tails. This decision is made by instructing a computer to generate a random number between 0 and 1. If this number lies on or between 0 and 0.5, then the value of heads is assigned. If this number is greater than 0.5, then the value of tails is assigned. In this way, a random number and an algorithm for its use can decide which value a specific variable is going to take. These values may then used in various computations to determine an experimental result.

At first glance, using random numbers to assign values to certain variables and then performing some calculations may not seem very experimental. However, biologists who were using Monte Carlo experiments in the 1960s considered them to be genuine experiments (Levin 1969, p. 38; Schull 1969, p. 47). Indeed, in his account of the development of the Monte Carlo technique

in physics, Peter Galison demonstrates how the practice of the Monte Carlo user closely mirrored the practice of traditional experimenters. Computer and non-computer experimenters shared the common concerns of tracking error, establishing replicability, and creating stability (Galison 1994). Galison's account of the experimental practice of Monte Carlo users in physics is rich and persuasive, but the case for considering Monte Carlo experiments as genuine experiments can also be made by comparing them to ordinary, non-computer experiments.[9] Moreover, comparing Monte Carlo and non-computer experiments reveals important structural similarities as well as important differences in the degree of experimental control between these two types of experiment.

In most non-computer experiments, an experimenter is trying to understand different features of a natural system. In many cases, this understanding is gained by manipulating or controlling certain elements of the system while allowing others to vary. The experimenter can then detect covariation among different elements of the system and even infer causal relationships between different elements (Cook and Campbell 1979, pp. 4–5). For example, suppose an experimenter is interested in the effects of temperature on *Drosophila* development. Specifically, suppose she is interested in how varying temperature during development will affect adult wing morphology. In setting up this experiment a number of factors are held constant or assumed not to affect the outcome. So, for instance, the diet, population density, or genetic background might all be things which the experimenter tries to keep constant across the groups of *Drosophila* under study. What is deliberately not kept constant is temperature. In this experiment, wing morphology is the outcome and is called the dependent variable, because it is dependent on a number of things most notably, the experimenter hopes, the independent variable or input. The independent variable or input is regarded as taking values which are not directly dependent on the other elements of the experiment. Temperature exposure, for instance, is not constrained by the *Drosophilas'* diet. A basic experimental setup then will have independent and dependent variables and a set of controlled parameters.

A key feature of this situation is the way the system is controlled to allow specific inferences about the relations between the dependent and independent variables to be made reliably. Inferences about the covariation of variables are usually made by comparing groups that receive different treatments; groups that have different values for the independent variable. In the case of the effect of temperature on wing morphology, for example, one group might be maintained at room temperature while another might be treated with a sudden drop in temperature twelve hours after fertilization. If the room temperature group is regarded as the norm or baseline, then the experiment would be to

determine the effects of the sudden drop in temperature. Variation in wing morphology can thus be understood in light of variation in temperature during development, other things being equal or controlled.

There are at least two senses of control in an ordinary experiment. On the one hand, the experimenter will want to control the environment in which the experiment takes place (Cook and Campbell 1979, p. 6). The idea is to keep extraneous forces from interfering with the outcome of the experiment. For experiments done in a field setting this can be a significant problem. On the other hand, the experimenter will want to control the effects of the independent variable. This type of control amounts to the ability to isolate the effects of the independent variable apart from the effects of any other element that might be correlated with it. Achieving these kinds of control is a way of eliminating threats to the validity of the inference of correlation between the dependent and independent variables (Cook and Campbell 1979, p. 8).[10]

Monte Carlo experiments share this basic structure of independent variables, dependent variables, and controlled parameters. The major difference between Monte Carlo experiments and ordinary experiments is that Monte Carlo experiments are experiments in a much more controlled environment. In Monte Carlo experiments the entire situation is computerized; creating a very contained and controllable environment. There is very little danger of spontaneous changes in the computer or computer program while they are running. Unfortunately, there are no such guarantees for experiments running in a complex natural environment. In virtue of their use of pseudo-random or quasi-random numbers as well as their computer implementation, Monte Carlo experiments do have characteristic kinds of error that have to be controlled. Huge literatures exist on variance reduction techniques and the evaluation of random numbers. These are not the same sources of error as those that may be faced in a statistical field experiment. Moreover, the kinds of error associated with Monte Carlo experiments are well known and can be readily addressed. The kinds of error associated with statistical field experiments are much more diverse, may not be easily detected, and may be very difficult to address and control. The major difference between Monte Carlo experiments and statistical field experiments, then, is in the number and diversity of kinds of error that need to be controlled. Monte Carlo experiments will in most cases be more controllable than will statistical field experiments, because they usually have fewer sources of error and fewer kinds of potential error.[11]

One of the purposes of an ordinary experiment is to increase our understanding of a natural system by carefully examining how systematic variation of certain elements affects other elements. For many geneticists in the late 1960s and early 1970s, Monte Carlo experiments were thought to be much

the same as ordinary experiments. For instance, at a conference on computer applications in genetics, B.R. Levin analyzed the assets and liabilities of Monte Carlo simulations reasoning that "A simulation study is an experimental study; one attempts to draw inferences about the nature of a system by manipulating some components of the system" (Levin 1969, p. 38). The skill of the Monte Carlo experimenter is put to the test in knowing how to construct the experiment, knowing which variables to manipulate, and knowing how to manipulate them (Levin 1969, pp. 38–39). Monte Carlo experiments and ordinary experiments are both characterized by a definite structure with specific kinds of controls. These controls allow the experimental situation to be carefully manipulated and allow inferences about relations between different variables of that system to be made reliably. It is these features that weigh in most strongly for the consideration of Monte Carlo experiments as genuine experiments.

Although it does not affect their status as experiments, comparing diffusion equation models with Monte Carlo experiments represents a significant departure from more typical uses of Monte Carlo experiments. The aim of the comparison of the results of diffusion models and the results of Monte Carlo experiments is to evaluate the computational adequacy of the diffusion model.[12] Most biologists in the 1960s and 1970s used Monte Carlo experiments for a different purpose; namely, to generate predictions when mathematical models proved intractable. These computer-generated predictions were then compared with the results of ordinary experiments. So in this more typical use, the outcomes of Monte Carlo experiments were directly compared to the outcomes of ordinary experiments, not to the outcomes of another mathematical model (see Ohta and Kimura 1974). The aim in the more typical use of Monte Carlo methods was to generate predictions for empirical testing. In the case presented here, Monte Carlo experiments were used to generate a set of results but they were not used as empirical predictions but as computational results. The aim of these experiments is an evaluation of computational adequacy, not empirical adequacy.

Motoo Kimura's Monte Carlo experiments

Motoo Kimura constructed eleven different Monte Carlo experiments to estimate the values of the effective number of alleles, n_e, and the actual number of alleles, n_a, for a number of different situations (see Table 1). These experiments were based on two different computer programs. In the first, mutation was deterministic (a predetermined number of mutations are introduced in each generation) and all members contributed equally to the gene pool. This program was intended to simulate a monoecious population. The second

Table 1. Results of Monte Carlo experiments on numbers of neutral alleles (after Kimura 1968b, p. 258, Table 2). Mutation is deterministic in experiments 1 and 2, but is stochastic in experiments 3 through 11. The numbers in parentheses indicate the numbers of outputs from which n_a and n_e were computed.

Expt.	Population				Mutation	Output	Observed means		Diffusion approx.	
no.	N	N_e	M	F	rate		n_a	n_e	n_a	n_e
1	100	100	\	\	0.005	100–1200 (56)	9.68	3.13	8.61	3.00
2	500	500	\	\	0.001	200–1000 (21)	13.43	2.79	11.82	3.00
3	100	50	25	25	0.005	120–2100 (100)	6.05	2.07	5.30	2.00
4	100	100	50	50	0.005	120–1200 (100)	9.34	2.26	8.61	3.00
5	100	18	5	45	0.005	120–1300 (60)	4.12	1.38	2.74	1.36
6	100	100	50	50	0.005	100–1200 (23)	10.91	3.22	8.61	3.00
7	100	50	25	25	0.005	100–1200 (23)	5.52	1.93	5.30	2.00
8	50	5	25	25	0.01	40–400 (19)	9.32	3.67	7.23	3.00
9	100	50	25	25	0.01	50–500 (19)	10.42	3.13	8.61	3.00
10	200	200	100	100	0.01	140–1120 (50)	34.74	10.66	27.30	9.00
11	500	167	50	250	0.001	220–900 (18)	6.78	1.99	5.07	1.67

program used random mutation and simulated a dieocious population. In this program there were usually equal number of males and females and it was possible for the effective population size to be less than the actual population size.[13] In both programs random numbers were used to determine where mutations occurred and how gametes were sampled. In the eleven experiments, population size, effective population size, and mutation rate were systematically varied.

Randomly sampling a certain number of gametes from an existing population to form another population would proceed as follows in a Monte Carlo experiment. These experiments start with 2N alleles, where N is population size, and every allele is unique at the start and every mutation will produce a new allele. In the parent population, the frequencies of the gametes, A_1, A_2, ... A_{2N}, are f_1, f_2, ... f_{2N}, respectively. The key to the experiment is to use random numbers to randomly assign each of the gametes to form the next population a value – to say whether it is an A_1 or a A_2, etc. To do so, a random number with a value between 0 and 1 is generated and compared with the existing gamete frequencies. If the random number lies between 0 and f_1, then gamete A_1 is generated. If the random number lies between f_1 and the sum of f_1 and f_2, then gamete A_2 is generated, and so on.[14] Each random number generated is used to determine the type of one gamete to be

contributed to the next population. Each such determination is called a trial. The desired number of gametes are selected by simply repeating this process. This set of trials constitutes a single run of the Monte Carlo experiment.

Each trial in a Monte Carlo experiment is like taking one spin of a roulette wheel. Different spins of a roulette wheel are not likely to produce the same result, and neither are different trials of the same Monte Carlo experiment. In order to get highly accurate results from a Monte Carlo experiment, a large number of trials must be made and statistically analyzed (Hammersley and Handscomb 1964, p. 19). The result given for a single run of a Monte Carlo experiment, therefore, is usually going to represent some statistical measure of a large number of trials.

In Kimura's experiments, outputs, measures of n_a and n_e, were given at pre-assigned intervals after a set number of generations had occurred. The outputs were taken after a number of generations when the process had reached a steady-state – a stationary distribution. This is important because on the face of it, it looks like Kimura only made a single trial with each of his Monte Carlo experiments, but we know that he should make many trials for accurate results. In this case, however, Kimura did not need to repeat his experiments from 0 to 1200 generations because he was trying to obtain values for n_a and n_e at equilibrium where they should not change significantly. Multiple samples from a single stationary distribution are equivalent to making multiple runs of the same experiment.

The outcome of Kimura's comparisons was fairly good agreement between the values predicted by the diffusion equation model and the values generated by the Monte Carlo experiments, with the exception that n_a tended to be underestimated by the diffusion equation model (Kimura 1968b, p. 257). Kimura concluded that these results were "satisfactory for practical purposes" (Kimura 1968b, p. 265).

Defending diffusion models

Philosophers of science usually consider comparative evaluations between the predictions of a theoretical model and a set of experimental results under the rubric of theory testing. While Kimura was certainly making a comparative evaluation between theoretical predictions and experimental results, there are good reasons *against* considering this evaluation to be a test. Instead, Kimura's comparative evaluation should be considered as a means of evaluating what I will call the computational adequacy of a highly idealized model.

In order for a comparison between a theoretical prediction and an experimental result to be considered a test, there must be the possibility that there could fail to be significant agreement between the prediction and the result.[15]

In the case of the comparison of the diffusion model and the Monte Carlo model, the Monte Carlo model was thought to capture Markovian properties of the Wright-Fisher model exactly. Moreover, it was widely accepted that if certain parameters of the Wright-Fisher model were allowed to approach zero then the process it modeled could be treated as a diffusion process (Feller 1951; Karlin and McGregor 1964). What this means is that if one assumes that mutation rate, u, and population size, N, are small and of the same order of magnitude, as Kimura, Ewens, and many others did, then the results of the Wright-Fisher model and the diffusion model should differ only by an order of magnitude (Ewens 1964a, p. 4).[16] The question then is, within this order of magnitude, how close is the agreement between the results of the two models? This is not the kind of case envisioned by philosophers where either an instance or a counterinstance is computed (Glymour 1980). The diffusion model was known to be an approximation of the Wright-Fisher model. The production of a counter instance was not possible given the parameter values and the model of statistical testing assumed. The worst that could happen was that the diffusion model would produce a poor estimation of the values of the Wright-Fisher model.

While this case is not a test, it is important. Kimura and others wanted to be able to use the diffusion model in place of the more intractable Wright-Fisher model. Because they knew that the diffusion model was based on a number of idealizations, they needed to evaluate how well the results of the diffusion model agreed with the results of the Wright-Fisher model. If there was a significant difference between results of the diffusion model and the results of the Monte Carlo experiment, then the computational adequacy of the diffusion model would be called into question. The ability of a model, often despite severe idealizations, to produce results that agree with the results of the model held as a comparative standard is a matter of what I will call the computational adequacy of the model.

It is important to note that computational adequacy is a feature of the model, not a feature of the results. Sets of results are evaluated in terms of their agreement to a standard or to another set of results. A model is judged to be computationally adequate if it is able to produce results which agree well with a standard.

Assessing computational adequacy is especially important for highly idealized models. In the case at hand, the diffusion model was known to be based on patently false idealizing assumptions. The major difference between the diffusion model and the Wright-Fisher model was the assumption that discreet processes could be treated as continuous processes. Evaluating computational adequacy demonstrates that, although these assumptions may not be empirically supported, they can be used without jeopardizing the predictive

adequacy of the model of which they are a part. Judging that the diffusion model is computationally adequate relative to the Monte Carlo experiments supports the interchangeability of the diffusion model and the Markov chain model used in the Monte Carlo experiments and by extension the Wright-Fisher model, despite the idealizations necessary for the diffusion model. Interchangeability in the context of computational adequacy does not mean that the two models are equivalent in all respects, but that they have the ability to generate the same or very similar predictive outcomes.

The empirical impact of computational adequacy

Shortly after Motoo Kimura began to defend the neutral theory of molecular evolution, his predicted values for the effective number of alleles and the average number of alleles became the object of a number of empirical tests. In 1972, George Johnson used empirical results from enzyme loci in different species of *Drosophila* to argue against accuracy of the infinitely many alleles model's predictions for n_e and n_a (Johnson 1972). Tsuneyuki Yamazaki and Takeo Maruyama countered Johnson's claim with more data and G.C. Kirby and R.B. Halliday used even more data to argue that values of n_e and n_a could not be used to discriminate between hypotheses of neutrality and selection (Yamazaki and Maruyama 1973; Kirby and Halliday 1973). In the context of these contradictory empirical tests, Kimura's evaluation of the computational adequacy of the diffusion version of the infinitely many alleles model takes on additional significance.

These empirical tests used predicted values of n_a generated by the diffusion model. Kimura's earlier evaluation of the computational adequacy of the diffusion model supported its interchangeability with the Markov chain model used in the Monte Carlo experiments and the Wright-Fisher model. Although the diffusion model was known to be highly idealized with respect to the Markov chain models, its computational adequacy relative to them sanctioned the use of its predicted values in place of theirs.

It is important to note that while the diffusion model was computationally adequate relative to the Monte Carlo model, it is not the case that the diffusion model was as empirically adequate as the Monte Carlo model. Computational adequacy is not equivalent to empirical adequacy. The significance of computational adequacy for empirical adequacy is more apparent if we first accept Elisabeth Lloyd's general account of theory testing in evolutionary biology. According to Lloyd, the evaluation of the relationship between a model and data, which lies at the heart of empirical testing, is based on three things: the fit between the model's results and the data, independent tests of different aspects of the model (including its assumptions), and variety of evidence

(Lloyd 1988, p. 145). Philosophical accounts of theory testing often focus exclusively on the issue of fit between the model and the data. Judgments of goodness of fit refer specifically to the extent of the agreement between the theoretical prediction and the experimental result. It is a matter of the comparison of outcomes.

Lloyd argues that the empirical adequacy of a model cannot be restricted to matters of fit alone, because various parts of biological models are often based on assumptions that have empirical interpretations (Lloyd 1988, p. 147). For instance, assuming random mating or the absence of migration or the absence of linkage are common practices in the construction of models in population genetics. Because these assumptions can be empirically interpreted, directly testing the empirical warrant for these assumptions "makes a difference to the empirical adequacy of the model" (Lloyd 1988, p. 148). Empirical adequacy then is more than just a matter of good fit; it is also a matter of well supported assumptions.

Computational adequacy in contrast to empirical adequacy simply is just a matter of agreement between outcomes where neither of those outcomes are the result of ordinary empirical observation or experimentation. In the case at hand, the diffusion model was judged to be computationally adequate and so could be interchanged with the Wright-Fisher model to produce predictions of the numbers of alleles. Either model could then be used to generate a prediction that could be compared with empirically obtained results and both should have roughly the same goodness of fit to those results. This does not mean that the two models are of equal empirical adequacy. The idealizations used in the diffusion model affect its empirical adequacy. The idealizing assumptions in the diffusion model are not independently and empirically supported. As such they impugn the empirical adequacy of the model. In making these assumptions, population biologists are weighing the value of tractability and computational adequacy against overall empirical adequacy, and in the end favoring tractability, computational adequacy, and the limited empirical adequacy produced by goodness of fit alone.

Conclusion

Monte Carlo experiments revolutionized the numerical evaluation of mathematical models by making it much easier and much faster. The result was much greater confidence in the diffusion equation approach and models based upon it. Where Warren Ewens had once painstakingly used numerical analysis to evaluate diffusion models, Monte Carlo methods allowed him to devise experiments which could be used to evaluate diffusion models with ease (Ewens 1963, 1964a, 1966). With Motoo Kimura's advocacy of the neutral

theory, diffusion models and this kind of computational testing took on an added significance.

The evaluation of diffusion models with Monte Carlo experiments is distinctly different from the evaluation of diffusion models with empirical results. Monte Carlo experiments share the same structure and attention to careful manipulation and control that statistical field experiments do. The main difference between Monte Carlo experiments and ordinary experiments is that the systems Monte Carlo experiments are engaged with are usually much more controlled. In the case discussed above, the Monte Carlo experiments were not used to generate a prediction for comparison to empirical results, but were used for the purpose of producing a set of results to compare against those produced by a related diffusion model. The distinguishing feature of this kind of comparative evaluation is that it is not an empirical test and so is not used to evaluate empirical adequacy but is used to evaluate computational adequacy.

Establishing that the diffusion model is computationally adequate relative to the Monte Carlo model and the Wright-Fisher model argues for the interchangeability of the diffusion model and the Wright-Fisher model for the purposes of computing numbers of alleles. Establishing high computational adequacy thus supports the use of the more tractable diffusion model by demonstrating that its idealizations do not hinder its computational adequacy. So, while idealizations are a definite liability in terms of empirical adequacy, they are not necessarily a liability in terms of computational adequacy. In its own way then evaluating computational adequacy demonstrates that highly idealized models do not necessarily limit a model's ability to produce empirically accurate predictions; in fact, in some cases highly idealized models will be the best way to produce empirically accurate predictions. At the same time, the highly idealized nature of these models reinforces the claim that there is more to empirical adequacy than just goodness of fit.

Acknowledgements

I am indebted to a number of people for their support, commentary and criticism. John Beatty, Richard Burian, James Crow, Warren Ewens, Paula Findlen, John Gillespie, James Griesemer, Elisabeth Lloyd, Sam Mitchell, William Provine and Paul Teller shared their insight on various relevant topics. Helpful commentary was also provided by members of the audience at the 1993 meeting of the International Society for the History, Philosophy, and Social Studies of Biological Science after a version of this paper was presented. The anonymous reviewers for this journal also provided valuable comments and criticisms.

Notes

[1] Haldane admits that his own passion was for difference equations.

[2] An infinitely many alleles model is a model of mutation and the creation of new alleles that assumes that there are a large enough number of mutational possibilities in a genome that every mutation can be treated as a mutation to a new allele.

[3] Stochastic models are models of random processes. They usually produce results in the form of a probablity distribution rather than a single value, which is characteristic of the results of deterministic models (Lloyd 1988, p 30).

[4] Genetic load is a way of speaking about the fitness reducing effects of harmful alleles. It was introduced by Herman J. Muller in 1950. Kimura's argument is based on cost of selection arguments introduced by J.B.S. Haldane in 1937. See Dietrich 1994.

[5] A stationary distribution is a distribution that does not change over time and is not dependent on the initial frequency (Gale 1990, p. 304).

[6] Effective population size is a way of discussing the size of the breeding population, while population size refers to all the individuals in a population. In more technical terms, "the effective number of a population is defined as the size of an idealized poulation that would have the same amount of inbreeding or of random gene frequency drift as the population under consideration" (Kimura and Crow 1963, pp. 279–280).

[7] Monoecious populations are single-sex populations. Dioecious populations have two sexes.

[8] The numbers used in Monte Carlo experiments are not random numbers; they may be pseudo-random or quasi-random numbers. True random numbers must be produced by a random process and pass statistical tests for randomness. Pseudo-random numbers are produced by deterministic processes and show no significant departure from randomness when statistically tested. Quasi-random numbers are random with respect to those statistical properties of interest, but are not random with respect to all of the other statistical properties typically used to evaluate randomness (Hammersley and Handscomb 1964, pp. 25–27).

[9] By ordinary experiments, I have in mind statistical field experiments done on naturally occuring systems. I do not wish to imply that experimentation is monolithic or run of the mill by labelling some experiments ordinary.

[10] See Cook and Campbell 1979, Chapter 2 for a detailed discussion of the variety of threats to experimental inference.

[11] This feature contributes to the swift acceptance of the results of the comparison between the diffusion model and the Monte Carlo model. It in effect allows the top of the black box to be closed much more quickly.

[12] More will be said about this mode of evaluation below.

[13] In a population with N_m breeding males and N_f breeding females the effective population size is $N_e = (4N_mN_f)/(N_m+N_f)$.

[14] This example is a simplified version of a Monte Carlo procedure presented by W. Hill and A. Robertson (1966, p. 273). It is cited as the procedure used by Tomoko Ohta and Motoo Kimura in their paper on simulation studies (1974, p. 618).

[15] Although this sentiment is usually invoked by philosophers in relation to deductive testing situations and the comparison of instances and counterinstances, the sentiment can be extended to testing of probabilistic models. The scientists discussed here adopted a statistical model of testing that posited a specific range of results expected at a given level of confidence. Results that fell outside that range can be counted as counterinstances with the given level of confidence. In the case discussed above, the result should not have fall outside the range if certain parameters are small and a large enough sample is taken. Failures of agreement were taken to be failures of measurement or estimation.

[16] I am indebted to John Gillespie and Warren Ewens for bringing this to my attention.

References

Cook, T. and Campbell, D.: 1979, *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, Houghton Mifflin Company, Boston.

Crow, J.: 1989, 'Twenty-Five Years Ago in Genetics: The Infinite Allele Model', *Genetics* **121**, 631–634.

Dietrich, M.: 1994, 'The Origins of the Neutral Theory of Molecular Evolution', *Journal of the History of Biology* **27**, 21–59.

Ewens, W.: 1963, 'Numerical Results and Diffusion Approximations in a Genetic Process', *Biometrika* **50**, 241–250.

Ewens, W.: 1964a, 'Correcting Diffusion Approximations in Finite Genetic Models', Technical Report 4, Department of Mathematics, Stanford University, Stanford.

Ewens, W.: 1964b, 'The Maintenance of Alleles by Mutation', *Genetics* **50**, 891–898.

Ewens, W.: 1964c, 'The Pseudo-transient Distribution and Its Uses in Genetics', *Journal of Applied Probability* **1**, 141–156.

Ewens, W.: 1965, 'The Adequacy of the Diffusion Approximation to Certain Distributions in Genetics', *Biometrics* **21**, 386–394.

Ewens, W. and Ewens, P.: 1966, 'The Maintenance of Alleles by Mutation – Monte Carlo Results for Normal and Self-Sterility Populations', *Heredity* **21**, 371–378.

Feller, W.: 1951, 'Diffusion Processes in Genetics', in J. Neyman (ed.), *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley, pp. 227–246.

Fisher, R.A.: 1922, 'On the Dominance Ratio', *Proceedings of the Royal Society of Edinburgh* **42**, 321–341.

Gale, J.: 1990, *Theoretical Population Genetics*, Unwin Hyman, London.

Galison, P.: 1994, 'Artificial Reality', in P. Galison and D. Stump (eds.), *The Disunity of Science: Boundaries, Contexts, and Power*, Stanford University Press, Stanford.

Gillespie, J.: 1989, 'When Not To Use Diffusion Processes in Population Genetics', in Marcus Feldman (ed.), *Mathematical Evolutionary Theory*, Princeton University Press, Princeton, pp. 57–70.

Glymour, C.: 1980, *Theory and Evidence*, Princeton University Press, Princeton.

Haldane, J.B.S.: 1933, *Science and Human Life*, Harper and Brothers Publishers, New York.

Hammersley, J. and Handscomb, D.: 1964, *Monte Carlo Methods*, Chapman and Hall, New York.

Hill, W. and Robertson, A.: 1966, 'The Effect of Linkage on Limits to Artificial Selection', *Genetical Research* **8**, 269–294.

Johnson, G.: 1972, 'Evidence that Enzyme Polymorphisms are Not Selectively Neutral', *Nature, New Biology* **237**, 170–171.

Karlin, S. and McGregor, J.: 1964, 'On Some Stochastic Models in Genetics', in J. Gurland (ed.), *Stochastic Models in Medicine and Biology*, The University of Wisconsin Press, Madison, pp. 245–279.

Kimura, M.: 1955, 'Stochastic Processes and Distribution of Gene Frequencies Under Natural Selection', *Cold Spring Harbor Symposium on Quantitative Biology* **20**, 33–53.

Kimura, M.: 1964, 'Diffusion Models in Population Genetics', *Journal of Applied Probability* **1**, 177–232.

Kimura, M.: 1968a, 'Evolutionary Rate at the Molecular Level', *Nature* **217**, 624–626.

Kimura, M.: 1968b, 'Genetic Variability Maintained in a Finite Population Due to Mutational Production of Neutral and Nearly Neutral Isoalleles', *Genetical Research* **11**, 247–269.

Kimura, M.: 1985, 'Diffusion Models in Population Genetics with Special Reference to Fixation Time of Molecular Mutants Under Mutational Pressure', in T. Ohta and K. Aoki (eds.), *Population Genetics and Molecular Evolution*, Japan Scientific Society Press, Tokyo, pp. 19–39.

Kimura, M. and Crow, J.: 1963, 'The Measurement of Effective Population Number', *Evolution* **17**, 279–288.

Kimura, M. and Crow, J.: 1964, 'The Number of Alleles That Can Be Maintained in a Finite Population', *Genetics* **49**, 725–738.

Kirby, G. and Halliday, R.: 1973, 'Another View of Neutral Alleles in Natural Populations', *Nature* **241**, 463–464.

Levin, B.: 1969, 'Simulation of Genetic Systems', in Newton Morton (ed.), *Computer Applications in Genetics*, University of Hawaii Press, Honolulu, pp. 38–46.

Lloyd, E.: 1988, *The Structure and Confirmation of Evolutionary Theory*, Greenwood Press, Westport.

Ohta, T. and Kimura, M.: 1974, 'Simulation Studies on Electrophoretically Detectable Genetic Variability in a Finite Population', *Genetics* **76**, 615–624.

Schull, W.: 1969, 'Discussion on Monte Carlo Simulation', in Newton Morton (ed.), *Computer Applications in Genetics*, University of Hawaii Press, Honolulu, p. 47.

Watterson, G.: 1962, 'Some Theoretical Aspects of Diffusion Theory in Population Genetics', *Annals of Mathematical Statistics* **33**, 939–957.

Yamazaki, T. and Maruyama, T.: 1973, 'Evidence that Enzyme Polymorphisms are Selectively Neutral', *Nature, New Biology* **245**, 140–141.