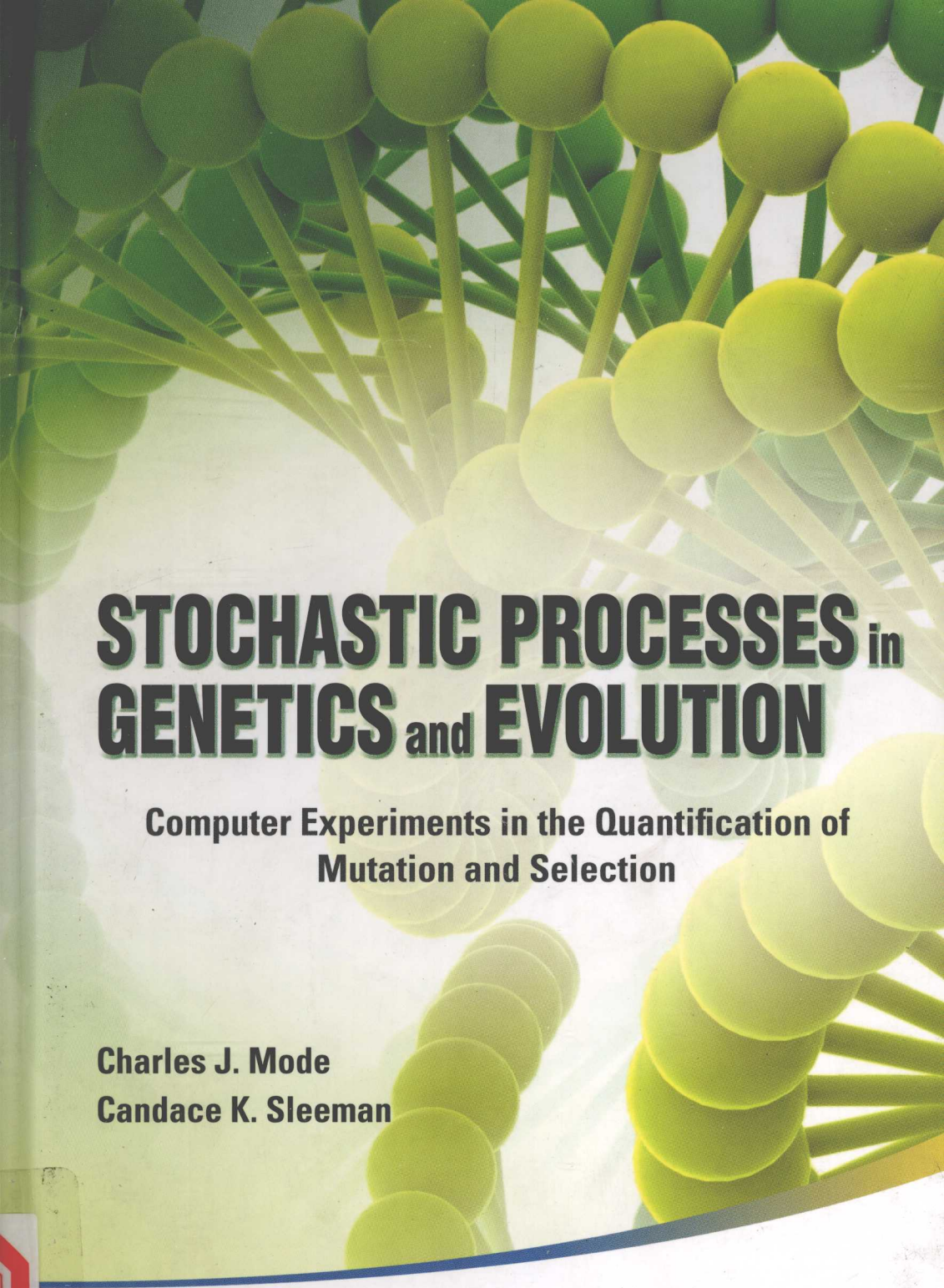

STOCHASTIC PROCESSES in GENETICS and EVOLUTION

Computer Experiments in the Quantification of Mutation and Selection

Charles J. Mode
Candace K. Sleeman

World Scientific

STOCHASTIC PROCESSES in GENETICS and EVOLUTION

Computer Experiments in the Quantification of Mutation and Selection

Charles J. Mode
Drexel University, USA

Candace K. Sleeman
NAVTEQ Corporation, USA

World Scientific

NEW JERSEY · LONDON · SINGAPORE · BEIJING · SHANGHAI · HONG KONG · TAIPEI · CHENNAI

Published by

World Scientific Publishing Co. Pte. Ltd.

5 Toh Tuck Link, Singapore 596224

USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601

UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library.

STOCHASTIC PROCESSES IN GENETICS AND EVOLUTION

Computer Experiments in the Quantification of Mutation and Selection

ISBN-13 978-981-4350-67-9

ISBN-10 981-4350-67-2

Printed in Singapore by Mainland Press Pte Ltd.

Prologue

At the outset it should be stated that this is not a book on phylogenetics in which millions of years of evolution and relationships among existing species are often under consideration. However, in chapters 6,7 and 8 stochastic models of nucleotide substitutions, which may be applied in research on phylogenetics, are reviewed and some useful extensions are suggested that accommodate nucleotide substitutions at a large number of sites of a DNA molecule rather than a single site or codons with three sites that are characteristic of most of the models introduced to the literature 3 to 4 decades ago. In contrast to research in phylogenetics, the main thrust of this book is to provide methods for simulating the stochastic evolution of single species during short periods of evolutionary time consisting of 10,000 to 200,000 years, but in some models time is expressed on a scale of generations.

A common theme of the computer simulation experiments reported in this book is the evolution of a population stemming from a small founder population. Of particular interest in these simulation experiments is the informative statistical summarization of a sample of Monte Carlo waiting times until a new beneficial mutation arises and become predominant in a population. Another distinguishing feature of this book is that all Monte Carlo simulation models are rooted in stochastic processes, and, in each case, an attempt has been made to present the mathematics in sufficient detail so that if an investigator were interested, he would, in principle, be possible to write software in a programming language of his choosing and duplicate any computer experiment reported in this book. The programming language used throughout this book was APL 2000, which is an international language that is popular among a minority of people who like the succinctness with which complex code may be written. Unfortunately, this programming language is not as popular as C^{++} and other languages. As

mathematics is a universal international language however, it is hoped that by inspecting the mathematics underlying a model, an investigator will be able to write software to implement any model discussed in a programming language of his choosing.

The following paragraphs of this prologue are devoted to suggestions that will be helpful to readers who wish to read this book thoroughly or merely skim through it or even skip some chapters to obtain an overall impression of the contents of the book. It is hoped that the modular nature of this book by topics will expedite this exploratory process.

Chapter 1 is devoted to an axiomatic treatment of probability, which will be useful in setting the stage for the chapters that follow. A central theme of this chapter is the concept of a finite probability space, which encompasses a sample space of outcomes of a conceptual experiment, a collection of events or subsets of the sample space and the definition a probability function defined on the class of events with certain properties. Random variables are then defined within the context of a probability space and the binomial, multinomial and Poisson distributions are derived and are applied extensively throughout the book. For those readers who are not comfortable with the axiomatic approach to probability, it will be sufficient to grasp the ideas underlying the binomial, multinomial and Poisson distribution. In this connection, a study of the many examples from Mendelian genetics involving applications of these distributions will be very helpful.

Chapter 2 is devoted the parameterization of the gametic distribution with respect to a large number of linked Mendelian loci or markers such as single nucleotide polymorphisms, $SNPs$, on molecules of DNA. After some suggestions for assembling data bases to study genetic recombination at the molecular level, a method is developed for parameterizing the gametic distribution in terms of recombination probabilities for some arbitrary number $N \geq 2$ of linked loci. In the closing section of this chapter, suggestions are made as to how the ideas developed in the foregoing sections can be applied to pedigrees in which linked markers are under consideration.

Chapter 3 is devoted large random mating diploid populations with no mutation or selection and the principal objective of the chapter is to develop a mathematical structure that can accommodate a large number of linked loci with a finite but arbitrary number of alleles at each locus so that convergence to a linkage equilibrium in such a population may be studied. This chapter begins with a classical account of convergence to linkage equilibrium for the case of two loci with two alleles at each locus, which may be found in many text books of population genetics. This result

is then extended to the case of a finite but arbitrary number of alleles at each locus and then finally to the general case of multiple loci mentioned above. Much of the content of chapter 3 is based on results by H. Geirenger, which were published in 1944. Among other things, this theory is based on a elegant application of set theory, and if this is discomforting to a reader, he can rest with the knowledge that this theory will not be used in subsequent chapters of the book. However, when genetic recombination is again encountered in chapter 14, the case of two linked loci discussed in chapters 2 and 3 will be applied to the case of two linked markers at the molecular level.

Chapter 4 is devoted to a presentation the Wright-Fisher process within the context of finite absorbing Markov chains in which applications of matrix theory are useful in reducing the structure of this process to simple terms that are familiar to anyone with a working knowledge of the theory of finite matrices. In particular, formulas of a set of conditional absorption probabilities are derived such that if the process starts in given transient state, the conditional probability that it is absorbed or the process in terminated in some particular absorbing state are expressed as elements of matrices. Furthermore, giving that the process terminates in some absorbing state, formulas for the conditional expectations and variances of the waiting times to absorption are derived. A general formula for the quasi-stationary distribution of a finite absorbing Markov chain are also derived, which will be useful in connections with branching processes introduced in subsequent chapters. Any mention of diffusion approximations to Wright-Fisher process have deliberately been avoided, because, for the most part, this book is devoted to computer intensive methods. In this chapter, Wright-Fisher processes with respect to a single autosomal locus with two alleles are the principal foci of attention and both the neutral case and the cases of mutation and selection as characterized within the Wright-Fisher paradigm in terms of probabilities. A class of Wright-Fisher processes with a state space such that all states communicate with each other was also included in this chapter.

Chapter 5 is devoted for the most part to Wright-Fisher process with multiple alleles at a single autosomal locus. As through trial and error it was found that matrix formulas derived in chapter 4 tend to become numerically unstable when the size of a Markov transition matrix exceeds about 1000×1000, it became necessary to use Monte Carlo simulation methods for dealing with process based on multiple alleles which usually entails the use of very large transition matrices. Fortunately, by using Monte Carlo

simulation methods, problems with numerical instabilities can be avoided and the evolution of populations consisting one million or more individuals may be studied for thousands of generations on desk top computers, where the execution times for each experiment with 100 or more replications may be accomplished within ten to twenty minutes. The application of Monte Carlo simulation methods in this chapter also necessitated a treatment of some theories underlying the computer generation of random numbers as well as the description of statistical methods for the informative summarization of Monte Carlo simulation data, which were subsequently used in some of following chapters of this book.

Chapters 6, 7 and 8 are devoted to the mutational process of nucleotide substitutions. In chapter 6, an overview of the fundamentals underlying Markov jump processes in continuous time with finite state spaces is given, and there is also a brief discussion of a probability space on which this class of stochastic processes lives. It is also shown that the exponential matrix function, provides a solution to the Kolmogorov differential equations for the case of finite state spaces in a general case. Following this overview, specific examples of nucleotide substitution models based on this class of processes are reviewed. For the more simple models, a symbolic computation engine was used to derive explicit formulas for the exponential matrix, which provides explicit functions for the matrix of current state probabilities of the process. For more complex examples, numerical forms of the exponential matrix were computed, given assumed numerical values of the parameters. In this chapter there are also discussions that illuminate the process of nucleotide substitutions if it does indeed follow the laws of evolution implicit in the theory of Markov jump processes in continuous time. For example, given values of the rate parameters of the process, one could estimate the expectation of the time a particular nucleotide spends at one site of a DNA molecule until a transition to another nucleotide at this site occurs.

When one considers the problem of extending a nucleotide substitution model from one site of a DNA molecule to many sites, a simple and straight forward approach to the problem would be that of assuming nucleotide transitions among sites occur independently with the same rate matrix for each site. But, it has been proposed in the literature, that substitutions may occur at different rate among the sites of a molecule and that sites may not evolve independently in a probabilistic sense. When there are different rate parameters for each site, the problem of dealing with many parameters also arises. Consequently, it becomes necessary to devise a structure that cir-

cumvents problems of dealing with many parameters and at the same time to formulate a model such that evolution of substitutions among sites are dependent in some sense. Such a formulation was considered in chapter 7.

Briefly, in this approach it was assumed that the rate matrices at many sites were realizations of a stationary stochastic process, depending on only a few parameters and constructed on the basis of a consistent family of finite dimensional distribution functions, where consistency is defined in papers and books dealing with the foundations of probability. Given a realization of this process, it was assumed that substitutions at different sites were governed by conditionally independent Markov jump processes with four states. When averaged over realizations of the rate process to obtain unconditional distributions of the processes of among the sites, however, independence among the sites no longer holds. It also suggested in this chapter that the rate process could be constructed form simple Gaussian processes based on first or second order autoregressive processes. The range of the sample functions of these processes is the real line $(-\infty, \infty)$, but this set can always be mapped into the set of positive real numbers in $(0, \infty)$, which is the rate space for Markov jump processes in continuous time.

Chapter 8 is devoted to a software implementation of the stochastic structure developed in chapter 7 along with computer simulation experiments on nucleotide substitutions in the D loop of the human mitochondrial genome, which consists of 1,120 base sites. The classification of Haplo groups in existing human population of the world is based on about 22 $SNPs$ that occur mostly in the D loop. Given an initial DNA sequence consisting of 1,120 bases, a computer simulation experiment was run until 22 mutation were accumulated and each of these experiments were replicated 50 times. The evolutionary time taken to complete each replication varied, because the rates governing the Markov jump processes with random. When classifying Haplo groups, two complicating issues are often mentioned. One is the problem of back mutation and the other is the problem of parallel mutations. By focusing attention within each replication, it was possible to write software to estimate the frequency of back mutation, in which the initial nucleotide at a given site mutates to some other base and then back mutates to the initial base at that site. It was also possible to estimate the frequency of parallel mutations among the 50 replication that could be interpreted as separate evolving human populations. A mutation is said to be parallel if same mutation at a particular site occurs in more than one replication. Back mutations complicate the classification of Haplo groups, because individuals who ought to belong to a group because

of shared ancestry are excluded due to the absence of a mutation. Parallel mutations complicate the problem of identifying a descendant of a founder of population based on whether he or she has a particular mutation, which could have been present in the actual founder of the group or is, on the other hand, the descendant of an individual who migrated into the group with a parallel mutation. Although attention was focused only on the D loop of the human mitochondrial genome, it is now known that the structure implemented in chapter could easily be extended to the entire mitochondrial genome that consists of about 16,000 bases or even large regions of DNA with larger number of bases.

Up until chapter 9, no stochastic structure had been entertained within which a range of demographic factors could be taken into account in simulating the evolution of human and other populations. As a first step towards correcting this omission, an outline of the one-type Galton-Watson branching process was given in chapter 9, which evolves on a time scale of discrete time generations. One of the serious limitations of this process, as well a the class of branching processes in general, which has been recognized for a long time, is that there are only two possibilities; namely, either the population becomes extinct or it increases without bound. To correct this limitation, the idea of a self regulating branching process was introduced. In this formulation, it is supposed that the probability that an individual in one generation survives to produce offspring in the next generation is a function of total population size with a parameter that indicates the population size that must be reached before extensive mortality occurs. It should be stated that this formulation is not simply a rework of the famous deterministic logistic model of population growth that has attracted much attention, because for some points in its parameter space, iterates of its defining equation become chaotic.

One of the innovational aspects of this chapter is that by using a one-type self regulating branching processes to simulate genealogies, it is possible to estimate the distribution of the number of generations back in time to the most recent common ancestor of any two randomly selected individuals in some current generation. This approach is an alternative way of looking at the problem of coalescence, which is mentioned in this chapter. A second innovation of self regulating branching processes was that it was possible to embed a deteministic model in a stochastic process, and it was shown by examples that, at some points in its parameter space, iterates of its defining equation become chaotic. Moreover, in computer simulation experiments, it was possible to compare the performance of statistically

summarized sample of Monte Carlo realizations of the stochastic process with the chaotic embedded deterministic model, see chapter 9 for details.

In order for Darwinian selection to occur, a population must contain two or more types. Consequently, in chapter 10 is devoted experiments in the quantification of mutation and selection within a framework of self regulating multitype branching processes, which evolve on a discrete time scale expressed in terms of generations. In these experiments, two components of natural selection were taken into account in the formulation. One was a measure of reproductive success as expressed in terms of the expected number of total offspring contributed by each individual of a given type to the next generation of an evolving population. The other component was the ability of individuals of each type to survive to produce offspring of the next generation. This ability was characterized in terms of parameters, depending on type, which provided a threshold value such that when total population size exceeded this threshold, the probability that an individual survived to reproduce was reduced. Only three types or genotypes were considered in the experiments reported in this chapter, and, mutations among the types were described in terms probabilities of one type mutating to another type per generation. In these experiments, selection was quantified by assigning numerical values to expectations of total number of offspring contributed the next generation to member of each type as well as the threshold parameters in the survival function. Mutations were quantified by assigning numerical values to the probabilities governing mutations among the types in each generation. It was also possible for this class of branching processes, to embed deterministic vector-valued non-linear difference equations in the process, so that trajectories of the evolution of the population could be compared with trajectories based on statistically summarized Monte Carlo simulation data.

One of the most innovative experiments reported in this chapter was that if one type had a threshold parameter that was greater than that of other types and the measures of reproductive success of the types were equal, then it was shown that the type with the greatest threshold parameter eventually rose to predominance in the population within a few thousand generations even though this type had arisen by mutation during the evolution of a small founder population in which the beneficial mutation was not present. Interestingly, in this experiment it was possible to study the rise of this mutation in a stochastically evolving population as well as a population evolving according to the embedded deteministic model. As illustrated in the Monte Carlo simulation data, there were high

levels of stochasticity in the beginning generations during the emergence of the mutation that was not, as expected, present in the trajectories computed using the embedded deterministic model. If an investigator confines his attention to the deterministic model in the studying the evolution of a multitype population, the presence of high level of stochasticity that exists during the emergence of a mutation would be missed.

Diploid populations of humans as well as those of other species are comprised of two sexes, females and males, who form sexual partnerships which usually contribute offspring to the next generation. When formulating stochastic models describing the evolution of such populations, it is necessary to include a module that can accommodate mating systems and the possibility that sexual selection may occur during mating process. Chapter 11 is devoted to the formulation and the computer implementation of stochastic models accommodating two sexes in populations that evolve in discrete time generations. In this class of models, the components of natural selection include the type of mating system, random or non-random, preferences of both females and males with respect to the phenotype in selecting their prospective sexual partners, measures of reproductive success by couple types as classified by genotype or phenotype and a probability that each individual female or male survives to contribute offspring to the next generation. Like all other stochastic population processes considered in this book, the stochastic process formulated in this chapter is also self-regulating. For the most part, the underlying genetics used in this chapter was confined to one autosomal locus with two alleles at each locus so that for both sexes only three genotypes were under consideration. Probabilities that each allele may mutate to the other per generation were also included in this formulation.

Perhaps one of most interesting computer experiments reported in this chapter was a case in which the only component of natural selection in force was a type of sexual selection in which females preferred only males of a certain genotype with a high probability. In an experiment in which only this type of selection was in force, it was shown that this sexual component of natural selection was sufficient for a mutant allele that produced the preferred genotype to drive this genotype to predominance in both sexes in a population that evolved from a small founder population in which the sexually preferred genotype was not present but had arisen from a mutation during the evolution of the two-sex population.

An obvious observation that may be made on any human population and many animal populations is that at any time the population consists of

overlapping generations corresponding to age cohorts in both females and males. It is thought by many that this overlapping of generation provided milieu for the passing on of culture from older to younger individuals and the evolution of high human cognitive abilities when compared to other animals. Chapter 12 is devoted to the development of a self-regulating two-sex stochastic population process that accommodates the evolution of an age structured population along with the genetics underlying such populations. From the vantage point of branching process, the formulation in chapter 12 is an algorithmic extension of what is known as the general branching process. Among the components of selection included in this age structured formulation are the parametrized risk functions of death by each age group in both the female and male populations. As of the complexity of this formulation, attention was confined to the embedded deterministic model for all computer experiments reported in this chapter, but as time passes attention will be devoted to problems centered around the development of algorithms to compute Monte Carlo realizations of the age structured process.

An ultimate goal of the research on the stochastic models of evolution presented in this book is to include the evolution of a genome at the molecular level and how processes of natural and artificial selected may have affected the structure of genomes. Before undertaking research to include models of genomes in stochastic models of evolution, it seems necessary to develop some acquaintance with recent literature on the concept of a gene. Accordingly, chapter 13 is devoted a review of biological literature on the history of the concept of gene and a peek at one of the most recent definitions of a gene. Basically, a region of DNA that codes for either proteins or $RNAs$ are frequently composed of introns and exons, and during the process of transcription and subsequent processing of the products of transcription, exons are sliced together in alternative ways to produce proteins that are used in various stages of development of an individual. But, the process of transcription is not an end in itself but is evidently controlled by regulatory regions of genomic DNA where other chemical structures, that may be coded for by other genes, bind and turn the transcription process of a given gene on and off. Thus, in order to get a grip as to how mutation may affect a gene, one must take into account not only the coding region of DNA but also those regulatory regions of DNA that turn coding regions off and on. The union of such regions of genomic DNA may involve thousands or even millions of bases as described in actual biological examples of genes in the closing sections of chapter 13.

Chapter 14, which is a small book within a book, is organized around three themes. One theme concerns a review of recent literature on developing computer models designed to simulate the evolution of model genomes that may consists of as many of a million or more base pairs in the presence of such forces of evolution as mutation and selection. Among those models reviewed, the accounts as to how mutation, selection and genetic recombination were incorporated into their formulations were not given is sufficient detail to suit members of the community of mathematical sciences, who may wish to replicate their reported results.

A second theme was a review of new methods for detecting signals of evolution in human haplotype data. This very recent research involved a model of genomic evolution that ran backwards in time to some founder population that existed about 40,000 or so years ago. Moreover, distributions of scores, representing measure of selection, which were estimated from simulated genomic data produced by the backwards in time simulation model, were applied in existing human haplotype data to detect signals of selection in relatively small regions of a genome that heretofore had not been possible to detected.

A third theme involved a phylogenetic study of several mammalian species designed to detect signals of natural or artificial selection in protein coding genes in several mammalian species whose genomes had been sequenced. The methods used in this research involved, among other things, a evolutionary model of three letter codons related to those discussed in chapter 6. Briefly, maximum likelihood estimates of parameters in these models were used to construct indicators of signals of selection in protein coding regions of the species under consideration.

Finally, the fourth theme of this chapter was centered around the problem of constructing stochastic models to simulate various types of mutations, that occur at the genomic level, as well as genetic recombination that included crossing over and gene conversion with respect to two makers during a phase of meiosis in diploid population when chromatids are arranged along a spindle of a dividing cell. Similarly, the various types of mutations, whose mathematical structure was thoroughly documented, were assumed to occur during meiosis and in that phase when the DNA content of the cell is doubled by a process that may involve that errors during the DNA copying process. The types of mutation included in this chapter were nucleotide substitution, deletions and insertions, copy number changes in finite segments of DNA and inversions of segments of DNA within a chromosome. Both the models of genetic recombination and mutations are documented in

sufficient detail so a read can discern their details and either accept, modify or reject a proposed formulation. This process acceptance, modification or rejection in a time honored way to progress in the development of models that eventually find acceptance in a given field of science.

The book ends with chapter 15, which is devoted to a suggested agenda for continuing the research projects proposed in the preceding chapters and a short review of books and material from other media that were sources of inspiration while writing and thinking about the contents of this book.

Acknowledgments

A number of people have been helpful in the writing and assembling the material used in this book. Of particular note is Towfique Raj, who as a undergraduate student at Drexel University requested a course from the senior author in Mathematical Genetics in the winter quarter of 2004, and it was his enthusiasm for the subject that awakened dormant interest in Genetics that eventually led to the writing of this book. Presently, Towfique is on a Postdoctoral Research Fellow, Harvard Medical School, Brigham Women's Hospital, Boston, MA 02115, and was very influential in assisting the authors in obtaining references and graphs for some of the material presented in chapter 14. Conversations with Warren Ewens of the University of Pennsylvania were also helpful in that he suggested that more work needed to be done on the modelling of selection and he has also read drafts of some of the chapters and offered suggestions for improvements.

Several contacts with working geneticists and others interested in evolution were also made at a symposium on evolution that was held in May and early June of 2009 at Cold Spring Harbor Laboratory on Long Island, New York in connection with the celebration of the 200-th anniversary of Charles Darwin's birth. Among the contacts at this symposium was Mike Levine of the University of California, Berkeley, who called attention to the mouse sonic hedgehog gene that was regulated by an enhancer 1 megabase away and would thus be a good biological example in thinking about computer models of genes as indicated in chapter 13. David Shaw, Mouse Genome Informatics, The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609, was also helpful also obtaining more information on this gene. David Gasser and Douglas Epstein, Geneticists at the University of Pennsylvania, were also very helpful in obtaining further information of this gene as well as information on an ortholog of this gene in humans, which has been implicated in holoprosencephaly.

Thanks are also due Peter Olofsson, Department of Mathematics, Trinity University, who read and offered suggestions for improvements of chapter 9, which were incorporated into the present version of this chapter. Margaret Dominy, Science Information Services, Drexel University Library, was also very helpful in providing copies of papers in older volumes of scientific journals that had not been digitized and were thus not available on the internet and also copies of papers in more recent publications on the internet. A professional proof reader and friend, the late Robert (Bob) Grant, also proof read several chapter of the book, which resulted in the correction of many minor errors. Another friend, Scott Frizen, read chapter 1 with close attention to details and offered suggestions that were incorporated into the present version of chapter 1. It is also fitting to acknowledge that the internet search engine, Google, played a significant role in finding many of the papers and other material cited in this book. Special thanks are also due to MATHWORKS, who permitted the use MATLAB free of charge in producing most of the graphs in the book. And finally, warm thanks to George Pearson and John MacKendrick of MacKichan Software, Inc., who provided invaluable and timely assistance in using the technical word processor, Scientific Workplace, which was used to write and produce the book.

Contents