

# Mathematical Models in Biology, An Introduction

Version 02.12.31

Elizabeth S. Allman<sup>1</sup>  
John A. Rhodes<sup>2</sup>

1 Department of Mathematics and Statistics,  
University of Southern Maine  
2 Department of Mathematics,  
Bates College

©2002, Elizabeth S. Allman and John A. Rhodes

To J., R., and K.,  
*may reality live up to the model*

# Contents

Preface	v
Note on MATLAB	ix
Chapter 1. Dynamic Modeling with Difference Equations	1
1. The Malthusian Model	1
What is a difference equation?	4
2. Non-linear Models	9
Creating a non-linear model	9
Iterating the model	10
Cobwebbing	12
3. Analyzing Non-linear Models	17
Transients, equilibrium and stability	17
Linearization	18
Oscillations, bifurcations, and chaos	20
4. Variations on the Logistic Model	28
5. Comments on Discrete and Continuous Models	33
Chapter 2. Linear Models of Structured Populations	35
1. Linear Models and Matrix Algebra	35
Vectors and matrices	37
2. Projection Matrices for Structured Models	45
The identity matrix and matrix inverses	49
3. Eigenvectors and Eigenvalues	54
The use of eigenvectors	55
Asymptotic behavior	57
Complex numbers	60
4. Computing Eigenvectors and Eigenvalues	65
Computer methods of calculation	67
Chapter 3. Non-linear Models of Interactions	69
1. A Simple Predator-Prey Model	69
The phase plane	71
2. Equilibria of Multipopulation Models	77
Nullclines and the direction of orbits	79
3. Linearization and Stability	82
4. Positive and Negative Interactions	87
Competition	87
Immune system vs. infective agent	88
Mutualism	88

Chapter 4. Modeling Molecular Evolution	93
1. Background on DNA	93
2. An introduction to probability	96
Mutually exclusive events and sums of probabilities	98
Independent events and products of probabilities	100
3. Conditional probabilities	106
4. Matrix models of base substitution	113
Markov models	115
The Jukes-Cantor model	116
The Kimura models	119
5. Phylogenetic distances	126
The Jukes-Cantor distance	127
The Kimura distances	129
Additive and symmetric distances: Log-det	129
Chapter 5. Constructing Phylogenetic Trees	137
1. Phylogenetic Trees	138
Topological trees	139
Metric trees	141
2. Tree Construction: Distance Methods – Basics	145
Rooting a tree	151
3. Tree Construction: Distance Methods – Neighbor Joining	155
4. Tree Construction: Maximum Parsimony	161
5. Other Methods	168
6. Applications and Further Reading	170
Chapter 6. Genetics	175
1. Mendelian Genetics	175
2. Probability Distributions in Genetics	186
The binomial distribution and expected values	186
The $\chi^2$ distribution	190
3. Linkage	198
Sex-linked genes	198
Linked genes and genetic mapping	200
4. Gene Frequency in Populations	212
Random mating and Hardy-Weinberg equilibrium	213
Fitness and selection	214
Genetic drift	217
Chapter 7. Infectious Disease Modeling	225
1. Elementary Epidemic Models	226
The <i>SIR</i> model	227
2. Threshold Values and Critical Parameters	231
The severity and duration of epidemics	233
3. Variations on a theme	239
The <i>SI</i> and <i>SIS</i> models	239
Contact rate and contact number	240
Immunization strategies	241
4. Multiple Populations and Differentiated Infectivity	248

Chapter 8. Curve Fitting and Biological Modeling	253
1. Fitting Curves to Data	254
Semilog and log-log graphs	256
Measures of error	258
2. The Method of Least Squares	262
3. Polynomial curve fitting	270
Modeling the growth of AIDS	270
Appendix A. Basic Analysis of Numerical Data	277
1. The Meaning of a Measurement	277
2. Understanding Variable Data — Histograms and Distributions	279
3. Mean, Median, and Mode	283
4. The Spread of Data	285
5. Populations and Samples	288
6. Practice	289
Appendix B. For Further Reading	291
Appendix. References	293
Appendix. Index	295



## Preface

Interactions between the mathematical and biological sciences have been increasing rapidly in recent years. Both traditional topics, such as population and disease modeling, and new ones, such as those in genomics arising from the accumulation of DNA sequence data, have made biomathematics an exciting field. The best predictions of numerous individuals and committees have suggested that the area will continue to be one of great growth.

We believe these interactions should be felt at the undergraduate level. Mathematics students gain from seeing some of the interesting areas open to them, and biology students benefit from learning how mathematical tools might help them pursue their own interests. The image of biology as a non-mathematical science, which persists among many college students, does a great disservice to those who hold it. This text is an attempt to present some substantive topics in mathematical biology at the early undergraduate level. We hope it may motivate some to continue their mathematical studies beyond the level traditional for biology students.

The students we had in mind while writing it have a strong interest in biological science, and a mathematical background sufficient to study calculus. We do not assume any training in calculus or beyond, as our focus on modeling through difference equations enables us to keep prerequisites minimal. Mathematical topics ordinarily spread through a variety of mathematics courses are introduced as needed for modeling or the analysis of models.

Despite this organization, we are aware that many students will have had calculus, and perhaps other mathematics courses. We therefore have not hesitated to include comments and problems (all clearly marked) that may benefit those with additional background. Our own classes using this text have included a number of students with extensive mathematical backgrounds, and they have found plenty to learn. Much of the material is also appealing to students in other disciplines who are simply curious. We believe the text can be used productively in many ways, for both classes and independent study, and at many levels.

Our writing style is intentionally informal. We have not tried to offer definitive coverage of any topic, but rather draw students into an interesting field. In particular, we often only introduce certain models and leave their analysis to exercises.

Though this would be an inefficient way to give encyclopedic exposure to topics, we hope it leads to deeper understanding and questioning.

Because computer experimentation with models can be so informative, we have supplemented the text with a number of MATLAB programs. MATLAB's simple interface, its widespread availability in both professional and student versions, and its emphasis on numerical rather than symbolic computation, have made it well-suited to our goals. We suggest appropriate MATLAB commands within problems, so that effort spent teaching its syntax should be minimal. While the computer is a tool students should use, it is by no means a focus of the text.

In addition to many exercises, a variety of projects are included. These propose a topic of study, and suggest ways to investigate it, but they are all at least partially open-ended. Not only does this allow students to work at different levels, it also is more true to the reality of mathematical and scientific work.

Throughout the text are questions marked with '►'. These are intended as gentle prods to prevent passive reading. Answers should be relatively clear after a little reflection, or the issue will be discussed in the text afterwards. If you find such nagging annoying, please feel free to ignore them.

There is more material in the text than could be covered in a semester, offering instructors many options. The topics of Chapters 1, 2, 3, and 7 are perhaps the most standard for mathematical biology courses, covering population and disease models, both linear and non-linear. Chapters 4 and 5 offer students an introduction to newer topics of molecular evolution and phylogenetic tree construction that are both appealing and useful. Chapter 6, on genetics, provides a glimpse of another area where mathematics and biology have long been intertwined. Chapter 8 and the appendix give a brief introduction to the basic tools of curve fitting and statistics.

In terms of logical development, mathematical topics are introduced as they are needed in addressing biological topics. Chapter 1 introduces the concepts of dynamic modeling through one-variable difference equations, including the key notions of equilibria, linearization, and stability. Chapter 2 motivates matrix algebra and eigenvector analysis through two-variable linear models. These chapters are a basis for all that follows.

An introduction to probability appears in two sections of Chapter 4, in order to model molecular evolution, and is then extended in Chapter 6 for genetics applications. Chapter 5, which has an algorithmic flavor different from the rest of the text, depends in part on the distance formulas derived in Chapter 4. Chapter 8's treatment of infectious disease models naturally depends on Chapter 3's introduction to models of interacting populations.

The development of this course began in 1994, with support from a Hughes Foundation grant to Bates College. Within a few years, brief versions of a few chapters written by the second author had evolved. The first author supplemented these with additional chapters, with support provided by the American Association of University Women. After many additional joint revisions, the course notes reached a critical mass where publishing them for others to use was no longer frightening. A Phillips grant from Bates and a professional leave from the University of Southern Maine aided the completion.

We thank our many colleagues, particularly those in the biological sciences, who aided us over the years. Seri Rudolph, Karen Rasmussen, and Melinda Harder



all helped outline the initial course, and Karen provided additional consultations until the end. Many students helped, both as assistants and classroom guinea pigs, testing problems and text and asking many questions. A few who deserve special mention are Sarah Baxter, Michelle Bradford, Brad Cranston, Jamie McDowell, Christopher Hallward, and Troy Shurtleff. We also thank Cheryl McCormick for informal consultations.

Despite our best intentions, errors are sure to have slipped by us. Please let us know of any you find.

Elizabeth Allman  
eallman@maine.edu  
John Rhodes  
jrhodes@bates.edu  
Portland, Maine  
Turner, Maine



## Note on MATLAB

Many of the exercises and projects refer to the computer package MATLAB. Learning enough of the basic MATLAB commands to use it as a high-powered calculator is both simple and worthwhile. When the text requires more advanced commands for exercises, examples are generally given within the statements of the problems. In this way, facility with the software can be built gradually.

MATLAB is in fact a complete programming language with excellent graphical capabilities. We have taken advantage of these features to provide a few programs making investigating the models in this text easier for the MATLAB beginner. Both exercises and projects refer to some of the programs (called **m-files**) or data files (called **mat-files**) below.

The **m-files** have been written to minimize necessary background knowledge of MATLAB syntax. To run most of the **m-files** below, say **onpop.m**, be sure it is in your current MATLAB directory or path and type **onpop**. You will then be asked a series of questions about models and parameters. The command **help onpop** also provides a brief description of the program's function. Since *m-files* are text files, they can be read and modified by anyone interested.

Some of the **m-files** define functions, which take arguments. For instance, a command like **compseq(seq1,seq2)** runs the program **compseq.m** to compare the two DNA sequences **seq1** and **seq2**. Typing **help compseq** prints an explanation of the syntax of such a function.

A **mat-file** contains data that may only be accessed from within MATLAB. To load such a file, say **seqdata.mat**, type **load seqdata**. The names of any new variables this creates can be seen by then typing **who**, while values stored in those variables can be seen by typing the variable name.

Some data files have been given in the form of *m-files*, so that supporting comments and explanations could be saved with the data. For these, running the *m-file* creates variables, just as loading a *mat-file* would. The comments can be read with any editor.

The MATLAB files made available with the text are:

- **aidsdata.m** — contains data from the CDC on AIDS cases in the United States
- **cobweb.m**, **cobweb2.m** — produces cobweb diagram movies for iterations of a one-population model; the first program leaves all web lines that are drawn, while the second gradually erases them
- **compseq.m** — compares two DNA sequences, producing a frequency table of the number of sites with each of the possible base combinations

- **distances.m** — computes Jukes-Cantor, Kimura 2-parameter, and log-det (paralinear) distances between all pairs in a collection of DNA sequences
- **distJC.m**, **distK2.m**, **distLD.m** — computes Jukes-Cantor, Kimura 2-parameter, or log-det (paralinear) distance for one pair of sequences described by a frequency array of sites with each base combination
- **flhivdata.m** — contains DNA sequences of the envelope gene for HIV from the ‘Florida dentist case’
- **genemap.m** — simulates testcross data for a genetic mapping project, using either fly or mouse genes
- **genesim.m** — produces a time plot of allele frequency of a gene in a population of fixed size; relative fitness values for genotypes can be set to model natural selection
- **informative.m** — locates sites in aligned DNA sequences that are informative for the method of maximum parsimony
- **longterm.m** — draws a bifurcation diagram for a one-population model, showing long-term behavior as one parameter value varies
- **markovJC.m**, **markovK2.m** — produces a Markov matrix with of Jukes-Cantor or Kimura 2-parameter form with specified parameter values
- **mutate.m**, **mutatef.m** — simulates DNA sequence mutation according to a Markov model of base substitution; the second program is a function version of the first
- **nj.m** — performs the neighbor-joining algorithm to construct a tree from a distance array
- **onpop.m** — displays time plots of iterations of a one-population model
- **primatedata.m** — contains mitochondrial DNA sequences from 12 primates, as well as computed distances between them
- **seqdata.mat** — contains simulated DNA sequence data
- **seqgen.m** — generates DNA sequences with specified length and base distribution
- **sir.m** — displays iterations of an SIR epidemic model, including time and phase plane plots
- **twopop.m** — displays iterations of a two-population model, including time and phase plane plots

Of the above programs, **compseq**, **distances**, **distJC**, **distK2**, **distLD**, **informative**, **markovJC**, **markovK2**, **mutatef**, **nj**, and **seqgen** are functions requiring arguments.

# Index

- absolute value, 89
- Africa, Out of, 212, 221, 257
- agouti fur, 284
- AIDS, 257, 341, 412, 416
- albinism, 295
- Allee effect, 44
- allele, 268
  - codominant, 280, 322, 325
  - dominant, 267, 268
  - fixation of, 329
  - frequencies, 322
  - multiple, 280, 339
  - mutation, 326
  - partially dominant, 279
  - recessive, 267, 268
  - semidominant, 279
  - wildtype, 301
- autocatalytic model, 22
- autosome, *see also* chromosome
- base substitution, 144, 173
- bases, 142
- basic reproduction number, 351, 365
- Bateson, William, 301
- bifurcation diagram, 30
- bin size, 428
- binomial coefficients, 297
- blood type
  - ABO system, 280, 334
  - MN system, 322
- bootstrapping, 254
- brachydactyly, 279
- cannibalism, 135
- carrying capacity, 15
- Centers for Disease Control (CDC), 261, 341, 412
- centromere, 306
- chaos, 31–32, 48
- characteristic equation, 98
- chickenpox, 32, 344, 352, 365
- $\chi^2$ -statistic, 289
- chromatid, 306
- chromosome, 268, 301
  - autosome, 305
  - homologous, 306
  - sex, 302, 305
- cobweb diagram, 17–19, 110
- codominance, *see also* allele
- codon, 143, 177
- color blindness, 304, 316, 333
- combinations, 284, 297
- competition
  - contest, 44
  - model, 105, 131–132, 137
  - scramble, 43
- competitive exclusion, 137
- complex numbers, 89–90
  - absolute value of, 89
- contact number, 365
  - maximal male and female, 380
- contact rate, 364, 366
- crossing over, 269, 305, 309, 311
  - interference, 320
- curve fitting, 385
  - least squares, 386, 399
  - line, 399, 406
  - polynomial, 412
- cystic fibrosis, 325
- deletion, 144
- Demography, Fundamental Theorem of, 87
- density dependence, 13
- determinant, 74, 97, 206–208
  - and inverse of matrix, 75
- deviation
  - mean, 438
  - mean square, 438
  - standard, 436, 438, 439, 441
  - total (TD), 394
- difference equation(s), 3, 5–6, 47
  - coupled, 50, 106
  - vs.* differential equations, 10, 47, 346
- differential equation(s), 10, 46, 47, 346
  - logistic, 48
- diffusion, 35
- diploid, 269, 305
- disjoint, 151
- distance
  - additive and symmetric, 199–202, 208, 217
  - genetic, 307, 308, 314
  - Jukes-Cantor, 195–198, 217, 222
  - Kimura, 198–199, 205–206, 222
  - linkage, *see also* distance, genetic

- log-det, 199–202, 206–208, 222
- methods of tree construction, 222
- phylogenetic, 142, 194–210
- physical, 308
- distribution, *see also* random variable, 428
- bimodal, 429, 432
- binomial, 282–288
  - expected value, 288, 298
- central tendency, 432
- $\chi^2$ , 288–293, 300
- continuous, 430
- discrete, 430
- normal, 428, 430, 438, 439
- probability, 282
- skewed, 429
- uniform, 429
- DNA, 141–145
  - aligned sequences, 145
  - coding, 143, 173
  - junk, 143
  - mutation, 144–145, 173
- dominance, *see also* allele
- Drosophila melanogaster*, 301
- edge, 213
- eigenvalue and eigenvector, 80–103, 178, 182, 206
  - complex, 89, 127
  - computation of, 97–103
  - dominant, 85
  - power method, 100
  - strictly dominant, 86
- emigration, 9
- equilibrium, 8, 23–24, 52, 80, 110, 116
  - saddle, 127
  - stable, 24, 112, 126
  - unstable, 24, 126
- Euler’s method, 47
- event(s), 148
  - complementary, 152, 157
  - independent, 154
    - definition of, 166
  - mutually exclusive, 150, 161
- expected value, *see also* random variable
- exponential model, 6
- extrapolation, 387
- fecundity, 3, 67
- $F_i$ , 267
- Fick’s law, 35
- Fitch-Margoliash
  - algorithm, 225–230, 235
  - method, 232–233
- fitness, 327
  - mean, 337
  - relative, 327
- fixed point, *see also* equilibrium
- Florida dentist AIDS cluster, 257, 261
- 4-point condition, 237
- fragile X syndrome, 304
- gametes, 268
  - random union of, 272
- GenBank, 257
- gene, 143, 265, 268
  - linkage, 304–314
    - cis and trans configurations, 318
  - polymorphic, 339
  - sex-linked, 301–304
- gene transfer, lateral, 213, 257
- genetic code, 143
- genetic drift, 330–333
- genotype, 268
  - parental type, 306
  - recombinant, 305–307
- geometric model, 6
- gonorrhea, 362, 376
- growth rate
  - finite, 3, 12
  - finite intrinsic, 15
  - intrinsic, 12, 86
  - per capita, 13
  - relative, 44
- haploid, 305
- Hardy-Weinberg equilibrium, 324
- hemizygote, 303
- hemophilia, 304, 315
- herd immunity, 366
- heterozygosity, 337
- heterozygote, 269
  - advantage, 330, 335, 340
- histogram, 427
- HIV, 257, 376
- hominoid, 211, 212, 256, 258
- homozygote, 268
  - advantage, 330, 335
- Huntington disease, 296
- hypothesis test, 289
- immigration, 9
- immune system model, 132–133
- immunization, 341, 349, 366
- independent assortment
  - of chromosomes, 305
  - of genes, 270, 272, 273, 277, 304, 307
- infectious disease
  - endemic, 360, 363, 379
  - epidemic, 341, 344, 350
- model
  - differentiated infectivity, 376
  - MSEIR, 372
  - SI, 362, 421
  - SIR, 344, 421
  - sir, 364
  - SIRS, 375
  - SIS, 362, 421
  - sexually transmitted (STD), 376

- infective class, 343
- influenza, 345
- informative site, 248
- inheritance
  - chromosomal theory, 305
  - Mendelian model, 268
- initial condition, 3
- insertion, 144
- interpolation, 387
- interquartile range, 436
- intersection, 153
- inversion, 144
- Jacobian matrix, 129
- Jukes-Cantor model, 217, 222
- leaf, 213
- least squares, *see also* curve fitting
- leprosy, 363
- lice, head, 345, 362
- likelihood, 253
- linear algebra, 53, 75
- linearization, 24–28, 124–126
- logistic model, 14, 28–32, 40, 106
- malaria, 347
- Malthus, Thomas, 6
- map
  - genetic, 307, 308, 313
  - linkage, *see also* map, genetic
  - physical, 314
- Markov
  - matrix, 178
  - model, 71, 176
- mass action, 107, 132, 345
- mating
  - assortative, 326
  - random, 323
- matrix
  - addition, 60
  - characteristic equation of, 98
  - definition, 54
  - identity, 72
  - inverse, 73
    - and determinant, 75
    - formula, 74
  - multiplication, 55, 57–59
  - projection, 54
  - scalar multiple, 60
  - singular, 75
  - transition, 54, 176, 177
  - transpose, 207, 404
- Maximum Likelihood method, 253, 254
- Maximum Parsimony method, 243–248, 253
  - assumptions of, 248
- mean, 433, 439, 440
- mean infectious period, 352
  - death adjusted, 374
- measles, 32, 345, 368, 373
- median, 433
- meiosis, 268
- meiotic drive, 340
- Mendel, Gregor, 266
- mitochondria, 212, 221, 256
- mixing, homogeneous, 107, 343, 345, 416
- mode, 432
- model
  - linear, 6, 51
  - non-linear, 13, 108
- molecular clock, 180, 197, 217, 225
- molecular evolution, 142
  - model, 173–193, 217, 253
    - equilibrium base distribution, 181
    - general Markov, 185, 199–201
    - Jukes-Cantor, 179–184, 194–198
    - Kimura, 184–185, 198–199
    - protein, 189
- mononucleosis, 363
- Morgan, Thomas Hunt, 301
- multinomial coefficients, 334
- mumps, 32, 373
- mutation, 141, 144
  - back, 145, 180
  - hidden, 145
- mutation-selection balance, 340
- mutualism model, 105, 133–135, 138–139
- Neighbor Joining algorithm, 235–238, 254
- normal equations, 404, 407, 414
- nucleotides, 142
- nullclines, 118, 120
- operational taxonomic unit (OTU), 213
- orbit, 111
- orthologous sequences, 212
- outgroup, 230, 242, 245
- overdominance, 330
- parallel evolution, 256
- parasites, 256
- parsimony score, 243
- partial derivative, 126
- pattern, 250
- pedigree, 278
- permutations, 297
- perturbation, 24, 124
- pertussis, 366
- phase plane, 110
- phenotype, 269
- physiology models, 47
- plot
  - log-log, 392
  - semilog, 391
- population genetics, 322–340
- population model
  - density dependent, 13, 40–46
  - discrete *vs.* continuous, 47–48
  - discrete logistic, 14

- harvesting, 36–37
- interacting, 105–139
- Leslie, 65–68, 89, 186
- linear, 6, 50–96, 385
  - intrinsic growth rate, 86
  - stable age/stage distribution, 87
- Malthusian, 2–14
- Markov, 71
- non-linear, 13–46, 105–139
- Ricker, 41, 45
- structured, 49
- Usher, 68, 136
- power method, 100
- predator/prey model, 105–130
- primate, 212, 258
- probability, 146–172
  - addition rule, 149, 151, 156–158, 161
  - conditional, 163–167
    - definition of, 165
  - frequency interpretation, 146
  - multiplication rule, 155–158
- Punnett square, 269–271, 274
- purine, 142, 144, 157
- pyrimidine, 142, 144, 157
- quarantine, 348, 361, 366, 371
- quartiles, 436
- rabies, 371
- random variable, 282
  - expected value, 287
    - additive property of, 288, 298, 308
- recessive, *see also* allele
- recombination frequency, 314
- regression, 407
- removal rate, 345, 352, 366, 377
  - relative, 351, 354, 379
- removed class, 343
- RNA, 143
- root, 213
- Rosco, 180
- rubella, 368, 373
- sample, 440
- scalar, 60
- segregation
  - of chromosomes, 268, 303, 305
  - of genes, 268, 270
- selection, 326, 330, 335
  - coefficient, 327
  - frequency-dependent, 340
- sensitivity, 168–169
- sensitivity analysis, 94
- sickle-cell anemia, 278
- significance level, 291
- smallpox, 342, 366
- specificity, 168–169
- spruce budworm, 46
- stability, 24, 110
  - analysis, 24–28, 124–128
    - by calculus, 26, 129
    - local *vs.* global, 28, 128
- stable age/stage distribution, 87
- statistics, 423
- steady state, *see also* equilibrium
- Stirling’s formula, 221
- Strong Ergodic Theorem, 87, 100–103, 186
- structurally unstable model, 113
- Sturtevant, Alfred, 307
- sum of squares for error (SSE), 394
- susceptible class, 343
- symbiosis, 134, 256
- syphilis, 362
- T-cells, 132
- taxon, 213
- Tay-Sachs disease, 274, 276, 282
- testcross, 276
  - three-point, 310
  - two-point, 308
- tetanus, 371
- tetrad, 306
- 3-point formulas, 226
- threshold value, 351, 367
- transient, 23, 116
- transition, 144, 157, 163, 184, 204
- transmission coefficient, 345, 363, 364, 377
- transversion, 144, 157, 163, 184, 204
- tree, 213
  - bifurcating, 213
  - construction
    - algorithms *vs.* optimality criteria, 253
    - methods, 222–255
  - metric, 216
  - neighbors, 235
  - parsimony score, 243
  - phylogenetic, 211, 212
    - applications of, 256
  - rooted, 214, 215
  - rooting, 230
  - topological, 214
    - number of, 216, 220–221
  - unrooted, 214, 215, 245
- tribolium*, 32, 135
- tuberculosis, 346, 363
- turbidity, 9
- union, 150
- UPGMA, 223–225, 229, 235
- vaccination, *see also* immunization
- variability in data, 255, 425, 436
- variance, 438
- vector
  - addition, 60
  - definition, 53
  - multiplication by matrix, 55
  - scalar multiple, 60



- vertex
  - interior, 213
  - terminal, 213
- whale hunting, 257
- yellow-lethal allele, 278, 296, 335
- zygote, 305