Introducing Transition Matrices and Their Biological Applications

Angela B. Shiflet
Department of Computer Science
Wofford College
Spartanburg, S. C. 29303 USA
001-864-597-4528
shifletab@wofford.edu

George W. Shiflet
Department of Biology
Wofford College
Spartanburg, S. C. 29303 USA
001-864-597-4625
shifletgw@wofford.edu

ABSTRACT

The Blue Waters Undergraduate Petascale Education Program (NSF) sponsors the development of educational modules that help students understand computational science and the importance of high performance computing. As part of this materials development initiative, we developed two modules, "Time after Time: Age- and Stage-Structured Models" and "Probable Cause: Modeling with Markov Chains," which develop application problems involving transition matrices and provide accompanying programs in a variety of systems (C/MPI, C, MATLAB, Mathematica). Age- and stage-structured models incorporate the probability of an animal passing from one age or stage to the next as well as the animal's average reproduction at each age or stage. Markov chain models are based on the probability of passing from one state to another. These educational materials follow naturally from another Blue Waters module, "Living Links: Applications of Matrix Operations to Population Studies," which provides a foundation for the use of matrix operations. This paper describes the two modules and details experiences using the resources in classes.

Categories and Subject Descriptors

K.3.2 [**Computers and Education**]: Computer and Information Science Education - Computer Science Education, Curriculum

General Terms

Design, Experimentation, Measurement.

Keywords

Computational Science, Matrices, Linear Algebra, Educational Modules, High-Performance Computing, Petascale, Blue Waters, Undergraduate.

1. INTRODUCTION

With NSF funding, the Blue Waters Undergraduate Petascale Education Program [1] is helping to prepare students and teachers to utilize high performance computing (HPC), particularly petascale computing, in computational science and engineering with the following three initiatives:

- Professional Development Workshops for undergraduate faculty
- Research Experiences for undergraduates
- Materials Development by undergraduate faculty for undergraduates

The goal of the Materials Development initiative is "to support undergraduate faculty in preparing a diverse community of students for petascale computing."

For this program, the authors developed and class tested the computational science related modules "Time after Time: Age- and Stage-Structured Models" and "Probable Cause: Modeling with Markov Chains," which are available at [2] and [3], respectively, on the UPEP Curriculum Modules site. This paper describes and discusses the modules and experiences using both in the course Modeling Biological Networks and class testing the first module in Linear Algebra and a course on Modeling and Simulation at Wofford College [4].

Several of the students in the classes at Wofford are pursuing an Emphasis in Computational Science (ECS). By taking Calculus I, Introduction to Programming and Problem Solving (in Python), Data Structures (in Python and C++), Modeling and Simulation, and Data and Visualization and doing a summer internship involving computation in the sciences, Bachelor of Science students may obtain an ECS [5]. Matrices are an important data structure in numerous computational models, and introducing transition matrices and eigenvalues with a variety of applications provides motivation to students in mathematics, computer science, and the other the sciences as well as in the Emphasis in Computational Science.

2. MODULES

2.1 Pedagogy

Prerequisites for the modules "Time after Time: Age- and Stage-Structured Models" and "Probable Cause: Modeling with Markov Chains" are minimal, requiring an understanding of matrix multiplication and the maturity to read the material but no programming or calculus background. Those who do not know how to multiply matrices or how to multiply a matrix times a vector might wish to cover first another Blue Waters module, "Living Links: Applications of Matrix Operations to Population Studies," by the same authors [6].

Students using the modules at Wofford College ranged from first- to fourth-year with majors from biology, chemistry, computer science, environmental studies, mathematics, physics, and undecided. The modules provide the biological background necessary to understand the applications; assuming an understanding of matrix multiplication, the mathematical background needed to complete the exercises and projects; and references for further study. Multi-part quick review questions throughout (three (3) in "Age- and Stage-Structured" and sixteen (16) in "Markov Chains") with answers at the end of the modules provide immediate feedback. The modules also have exercises (five and three, respectively) for reinforcement and practice and project assignments (eight or nine, respectively) for further exploration using a computational tool.

To aid in exploration of the multi-scale aspects of the science and the computing process, example solutions involving serial and parallel model development accompany the modules. For an age-structured model, serial programs are available in MATLAB, Mathematica, and C, while HPC programs in C with MPI illustrate parallel parameter sweeps and matrix partitioning. Bioinformatics programs using Markov models to help locate genes are available in MATLAB, C, and C/MPI. (Blue Waters Student Intern Jesse A. Hanley implemented a matrix partitioning program, and Intern Whitney E. Sanders developed the parameter sweeps and Markov model in C/MPI.) Several datasets for use in projects also accompany the modules.

2.2 Age- and Stage-Structured Matrices: Module Content and Applications

"Time after Time: Age- and Stage-Structured Models" considers situations that classify individuals in a species by age, such as Years 1, 2, and 3, or stage, such as larvae, juvenile, and adult. Solutions employ matrices to determine the intrinsic growth rates, the proportion of each group in a stable distribution, and how sensitive the long-term population growth rate and predicted time of extinction are to small changes in parameters. We can employ the latter to determine the best category to target for conservation efforts for endangered species and for eradication efforts for pests.

Figure 1 presents a state diagram for a problem with the states denoting ages (Year 1, 2, or 3) of a bird. The left-pointing arrows represent fecundity or reproduction: A Year 2 (ages 1-to-2 years old) mother has a mean of five (5) female offspring, while a Year 3 (ages 2-to-3 years old) mother has four (4) female offspring on the average. The right-pointing arrows indicate survival rates of $P_1 = 15\%$ and $P_2 = 50\%$ from Year 1 to Year 2 and from Year 2 to Year 3, respectively. The information can be consolidated into a matrix, called a Leslie matrix, as follows:

$$\begin{bmatrix} 0 & 5 & 4 \\ 0.15 & 0 & 0 \\ 0 & 0.50 & 0 \end{bmatrix}$$

The module shows that over time the percentage of eggs/chicks stabilizes to 82.06% of the total population, while Year 2 birds comprise 12.05% and Year 3 birds are 5.90% of the population. Moreover, eventually each age group changes by a factor of $\lambda = 1.0216$ (102.16%) from one year to the next, and this λ is the dominant eigenvalue for the matrix.

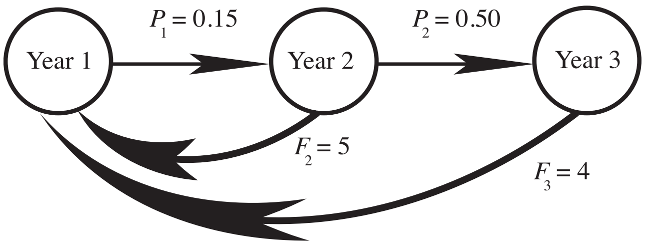

Figure 1. State diagram for age-structured problem

The sensitivity of λ with respect to P_i, $(\lambda_{new} - \lambda) / (P_{i,new} - P_i)$, measures the numeric impact on λ of a change in P_i. For small changes in P_i, the module shows that λ is most sensitive to changes in survivability of Year 1 birds, P_1. Thus, conservationists should probably concentrate their efforts on helping eggs and nestlings survive.

The module also covers a stage-structured model of the Indo-Pacific lionfish, an invasive and destructive species to reef habitats. Figure 2 illustrates that the model also includes probabilities for an animal remaining at the juvenile and adult stages. From this information, we can form a matrix similar to the Leslie matrix, called a Lefkovitch matrix, as follows:

$$\begin{bmatrix} 0 & 0 & 35315 \\ 0.00003 & 0.777 & 0 \\ 0 & 0.071 & 0.949 \end{bmatrix}$$

Module material and a project explore intrinsic grow rate, stable population distribution, and sensitivity analysis to make recommendations for controlling this menace.

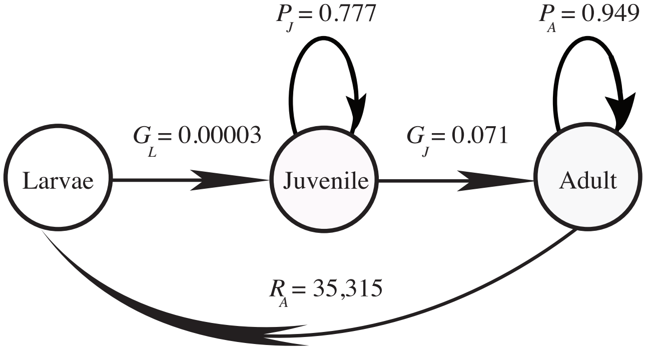

Figure 2. State diagram for stage-structured problem

2.3 Markov Chains: Module Content and Applications

The module "Probable Cause: Modeling with Markov Chains" also considers biological problems whose solutions involve transition matrices. Markov chain models (MCM) are based on the probability of passing from one state to another. Developing the necessary probability theory and biological background, the module solves a variety of problems using MCM from predicting the behavior of animals to locating genes in the DNA.

One problem considers a simplified system where the monkey is only in two states, eating (E) and resting/sleeping (R). Figure 3 enumerates the probabilities of moving between states, and the following transition matrix captures this data:

$$\begin{bmatrix} 0.6 & 0.2 \\ 0.4 & 0.8 \end{bmatrix}$$

As time passes, the module shows that the proportions approach 1/3 and 2/3 for eating and resting, respectively.

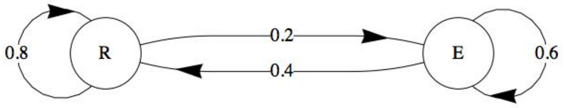

Figure 3. State diagram for Markov chain problem

We can employ such Markov models in a variety of problems in bioinformatics, which deals with the organization of biological data, such as in databases, and the analysis of such data. In a similar fashion to the example of changes of state for the monkey, a Markov chain can model the mutation process in DNA.

Moreover, the GeneMark algorithm employs Markov models to help locate genes in a DNA sequence. In many organisms, the sequence of bases CG appears less that we would expect from random occurrences of C and G independently. However, small regions, called CpG islands, upstream (before) of many genes are rich in the sequence CG; so we can employ CpG islands to locate genes. The module body develops a simplified 1st-order Markov model using the probabilities of bases and pairs of bases, while a project considers the more involved GeneMark 5th-order Markov algorithm employing probabilities involving quintets and sextets of bases.

2.4 High Performance Computing in Modules

Following the aims of UPEP, both modules have example programs and sections that focus on high performance computing (HPC) related to their particular applications. The section "Parameter Sweeping with High Performance Computing" in the age-structured module discusses the utility of parameter sweeping, or executing a model for each element in a set (often a large set) of parameters or of collections of parameters. As stated in the module, "The results can help the modeler obtain a better overall picture of the model's behavior, determine the relationships among the variables, find variables to which the model is most sensitive, find ranges where small variations in parameters cause large output changes, locate particular parameter values that satisfy certain criteria, and ascertain variables that might be eliminated to reduce model complexity" [7]. Besides being very useful, such parameter sweeping is embarrassingly parallel.

An algorithm for finding genes discussed in the Markov chains module is also embarrassingly parallel. Using a particular Markov model to score every subsequence of 200 bases, high scores indicate a greater likelihood that the subsequence is in a CpG island and that a gene is to follow. Multiple processes can evaluate scores for different subsequences, speeding the task significantly. Besides this specific example, a section on "High Performance Computing and Bioinformatics" discusses the utility of high-performance computing in a variety of other applications in bioinformatics.

2.5 Blue Waters UPEP Internship Involvement

During the summer of 2010 and following academic year, student Jesse Hanley held a Blue Waters UPEP Internship to develop parallel versions of programs using C and MPI to support "Age- and Stage-Structured Models" and other modules. The following year, Blue Waters intern Whitney Sanders continued Jesse's efforts with HPC programs for both modules discussed in this paper. Their programs accompany the modules on the NCSI UPEP Curriculum Modules site [2]. Jesse's is planning to work in the HPC field, and both students will be pursuing graduate work.

2.6 Exercises and Answers in Modules

After the body of educational material, each module contains a section with multi-part exercises, while a subsequent section has answers to selected parts. For example, the section in the "Age- and Stage-Structure Models" includes five exercises with one problem from the research literature involving loggerhead sea turtles.

2.7 Projects

After the exercises, each module presents eight or nine large projects for students to complete as individuals or with a team. Instructions indicate to develop sequential or high performance computing versions.

"Age- and Stage-Structure Models" has projects based on the research literature including ones on Uinta ground squirrels, skate, red-cockaded woodpeckers, lionfish, Pacific salmon, Furbish’s lousewort, and cane toads in Australia. An additional project involves determining and graphing the speedup factor versus the number of processes for various parameter sweeps.

The projects in the "Markov Chains" module include problems on the shapes of epithelium cells, succession in a forest, the dynamics of cattle fecal shedding of a pathogen, development of the BLAST algorithm for non-gapping local alignment of DNA segments, determination using the GeneMark algorithm of the most likely candidates for subsequences being in CpG islands, the Stepping Stone Model useful in genetics, and DNA sequence evolution.

3. TESTING AND EVALUATION

3.1 Class Testing of Age- and Stage-Structured Module

"Age- and Stage-Structured Models" was class tested in three courses, Modeling and Simulation in fall of 2011, Modeling Biological Networks in January of 2012, and two sections of Linear Algebra in spring of 2012. In the first class, two Emphasis in Computational Science (ECS) students successfully implemented a system dynamics model for the age-based example in the text.

Four biology majors and one triple major in chemistry, mathematics, and computer science considered the module in greater depth in a January interim on Modeling Biological Networks taught by the authors. During the interim, students take only one course not in the usual curriculum. Each day, students attended class for three hours, where they made presentations and worked on projects, and then continued developing projects between classes. Four of the students were freshmen, while one was a sophomore; and three of the students are pursuing the ECS. Before considering "Age- and Stage-Structured Models," the class worked through two MATLAB tutorials and the module "Living Links: Applications of Matrix Operations to Population Studies" to gain a background in matrices [6]. Students read each module and worked exercises before class and took a short quiz on the quick review questions in class. Over a three-day period, each student individually or with a partner developed and presented an age-structured and a stage-structured project. In all, the class successfully completed six different models using MATLAB.

Also reading the material before class, forty-one (41) students in two sections of Linear Algebra taught by Dr. Ted Monroe studied age-structured models one day the last week of classes in Spring, 2012. Students in this course, typically sophomores, come from a variety of majors and minors including mathematics, computer science, biology, chemistry, and physics. During class, the professor focused more on the mathematics and less on the biology for about 30 minutes. Examining the bird population diagram in Figure 1, he ensured that the students understood the system of equations, notation, and the matrix-vector equation. He reminded the class that the system is equivalent to a matrix-vector equation and that that they had looked at linear difference equations previously. Unlike problems where populations stabilized, he noted that this population problem led to growth over time. Also, the professor pointed out the population growth equations in annual and exponential form. However, the focus of the class was on the connection of the model to eigenvalues and eigenvectors.

3.2 Evaluation of Age- and Stage-Structured Module

Immediately after using the material, students in the interim on Modeling Biological networks completed a questionnaire about the module. The questionnaires had the students rate the following statements from 1 (strongly disagree) to 5 (strongly agree):

- I understood the science applications in the module.
- I understood the mathematics in the module.
- The module was readable.
- The Quick Review Questions helped me understand the material.
- The exercises helped me understand the material.

Means of the four responses were 4.5, 4.0, 4.25, 4.75, and 4.5, respectively.

In Linear Algebra, the professor had the students complete the questionnaire at the beginning and the end of class. Excluding students who had not read the material in the "before" category, Table 1 summarizes the results for questions 1-3. Unfortunately, the page with answers to the Quick Review Questions was not included with their materials, reducing their effectiveness. Moreover, students were not required to work exercises before class. Thus, particularly for the "Before" column, answers to the first three questions are more meaningful than those for the last two questions.

Table 1. Means, 1 (strongly disagree) to 5 (strongly agree)

Question	Before	After
I understood the science applications in the module.	4.22	4.45
I understood the mathematics in the module.	4.24	4.74
The module was readable.	4.16	4.40

Two elaborated on the above scores, "I enjoyed reading about applications," and "The mathematics in this module was easy to understand and the questions helped to reinforce what I had read."

Some of the comments in the questionnaire given at the beginning of class on what the student liked best about the module follow: "The module was easy and enjoyable to read. The information on the turtles was really interesting." "I liked how it was able to apply to real-life situations." "Instantly visualizable, elegantly simple, easy to understand what each structure/value represented." "I liked that the math we are learning now can be applied to a real-world problem to help understand endangered species." "I found it interesting that the example used actually converges to a specific percentage ratio." "Quick review questions and examples." "The introduction was very engaging and the mathematics was displayed very clearly and concisely." "I liked the subject matter and the mathematical applications." "I enjoyed the complete description of everything discussed and its relevance to science." "The module was very readable. The science applications were well explained, and the examples were helpful." "I liked that the module was easy to follow and the applications were clear." "The mathematical procedures were very easy to follow. I also liked the real-life applications of this kind of math." "The module was set up very well. The introduction and topic were interesting." "I thought the explanations of the science was interesting so it made following the math easier." "I enjoyed the idea of linking mathematical models to explain and possibly solve world issues I appreciate personally as an environmental enthusiast." "I liked the science applications. I'm a science person in general, and I also enjoy math, so putting the two together makes me happy." "The projected population growth rate [using the dominant eigenvalue] was very interesting." "I liked the biology aspect of it! (I'm a biology major, and really like how math works with living systems.)"

The few students who indicated any difficulties with the module were challenged by the mathematics, particularly the concept of "eigenvalue." However, on the questionnaire at the end of class, they indicated they now understood the concept and how to compute the eigenvalues. One student found typographical errors in the module, which the authors have corrected.

3.3 Class Testing of Markov Chains Module

After completing the "Age- and Stage-Structured Models," for three days the Modeling Biological Networks interim class turned its attention to "Markov Chains." Reading the module before class and completing two exercises, the group checked the exercises in class and had a quiz on selected quick review questions. Students in pairs or individually developed and presented two models each with the class completing a total of five projects.

3.4 Evaluation of Markov Chains Module

With the same questionnaire as for "Age- and Stage-Structured Models," averages on the questions (4 responses) were 5.0, 4.0, 4.5, 4.5, and 4.25, indicating that the most challenge came from the mathematics in the module. Some responses to a question on what the student found most difficult in the module reinforced this perception: "Figuring out the math with the transition matrices and the length-normalized log-odds score," and "I didn’t fully understand some of things in the module until we talked about them in class the next day (i.e., transition matrices and ultimate distributions)."

However, other remarks on what the student liked best included the following: "It was easy to understand, and it gave everything to the reader that the reader needed to know." "Great explanations of calculating probabilities." "The definitions/rules sections always help me understand the material in these modules. Having answers to the QRQ's is also helpful for understanding the content." "I enjoyed the quantity of helpful exercise and quick review problems."

Some of the additional comments were as follows: "I have always found probabilities to be difficult, but this module helped me understand them much better." "I like that all of the material ties into real-life problems. It makes everything much more interesting and sometimes more understandable as well." "I found this module very helpful. I was fully able to quickly grasp the information provided in each section and use the knowledge to work the review questions." "This is a very well thought out, organized, and helpful module."

One suggestion was, "Include answers to the exercises." In response, the authors added a section of answers to selected exercise parts.

4. CONCLUSION

"Time after Time: Age- and Stage-Structured Models" and "Probable Cause: Modeling with Markov Chains" and their associated programs in MATLAB, Mathematica, and C/MPI are available on the UPEP Curriculum Modules website [2, 3]. Class testing of the modules in Modeling Biological Networks, Modeling and Simulation, and Linear Algebra helped refine the modules and showed their utility in introducing applications of matrices, eigenvalues, parameter sweeping, and HPC concepts. High questionnaire scores and enthusiastic comments from undergraduates in three different types of courses verify the conclusion that "Time after Time: Age- and Stage-Structured Models" and "Probable Cause: Modeling with Markov Chains" can be an effective educational modules in a variety of classes, levels, and settings.

5. REFERENCES

[1] National Computational Science Institute Undergraduate Petascale Education Program (UPEP). `http://computationalscience.org/upep` Accessed 3/5/11.

[2] Shiflet, A. and Shiflet, G. 2012. "Time after Time: Age- and Stage-Structured Models." National Computational Science Institute Undergraduate Petascale Education Program (UPEP) Curriculum Modules, UPEP Curriculum Modules site. `http://www.shodor.org/petascale/materials/UPModules/ageStructuredModels/` Accessed 5/21/12.

[3] Shiflet, A. and Shiflet, G. 2012. "Probable Cause: Modeling with Markov Chains." National Computational Science Institute Undergraduate Petascale Education Program (UPEP) Curriculum Modules, UPEP Curriculum Modules site. `http://shodor.org/petascale/materials/UPModules/probableCause/` Accessed 5/21/12.

[4] Wofford College. `http://www.wofford.edu/` Accessed 3/5/11.

[5] Computational Science - Wofford College. `http://www.wofford.edu/computationalscience/` Accessed 3/5/11.

[6] Shiflet, A. and Shiflet, G. 2011. "Living Links: Applications of Matrix Operations to Population Studies." National Computational Science Institute Undergraduate Petascale Education Program (UPEP) Curriculum Modules, UPEP Curriculum Modules site. `http://shodor.org/petascale/materials/UPModules/populationMatrices/` Accessed 5/21/11.

[7] Luke, Sean, Deeparka Sharma, Gabriel Catalin Balan. "Finding Interesting Things: Population-based Adaptive Parameter Sweeping." 2007. In GECCO '07: Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation. Pages 86-93. ACM.