Introducing Matrix Operations through 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
For the Blue Waters Undergraduate Petascale Education Program (NSF), we developed a computational science module, "Living Links: Applications of Matrix Operations to Population Studies," which introduces matrix operations using applications to population studies and provides accompanying programs in a variety of systems (C/MPI, MATLAB, Mathematica). The module provides a foundation for the use of matrix operations that are essential to modeling numerous computational science applications from population studies to social networks. This paper describes the module; details experiences using the material in two undergraduate courses (High Performance Computing and Linear Algebra) in 2010 and 2011 at Wofford College and two workshops for Ph.D. students at Monash University in Melbourne, Australia, in 2011; and describes refinements to the module based on suggestions in student and instructor evaluations.

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

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

The Materials Development initiative has as its goal "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 modules "Living Links: Applications of Matrix Operations to Population Studies," which is available at [2] on the UPEP Curriculum Modules site. This paper describes and discusses the module and our experiences using it in the courses High Performance Computing and Linear Algebra at Wofford College [3] in 2010 and 2011, respectively, and using some of the examples, applications, and projects from the module in "Quantitative Modelling Using MATLAB: Introduction," a component of two 2011 workshops ("Introduction to Computational Thinking" and "Computational Workshop for the Life Sciences: Bringing Computation to Life") for Ph.D. graduate students, sponsored by the Monash eScience and Grid Engineering Laboratory (MeSsAGE Lab) at Monash University in Melbourne, Australia [4].

Several of the students in the classes at Wofford are obtaining the Emphasis in Computational Science (ECS). Bachelor of Science students may obtain an 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 [5]. Matrices are an important data structure in numerous computational models, and introducing operations on matrices with population applications provides motivation to students in mathematics, computer science, and the other the sciences as well as in the Emphasis in Computational Science.

2. MODULE

2.1 Pedagogy

Prerequisites for the module "Living Links: Applications of Matrix Operations to Population Studies" are minimal, requiring the maturity to read the material but no programming or calculus background. Students who used the module at Wofford College ranged from first- to fourth-year with majors from biology, chemistry, physics, mathematics, computer science, and undecided. Those attending the workshops at Monash University were mainly Ph.D. science students from such areas as biology, chemistry, engineering, mathematics, psychology, pharmacy, and medicine. The module provides the biological background necessary to understand the applications, the mathematical background needed to complete the exercises and projects, and references for further study. Ten (10) multi-part quick review questions with answers at the end of the module provide immediate feedback. The module also provides seventy-five (75) exercises for reinforcement and practice and seven (7) project assignments for further exploration using a computational tool.

To help with implementation of models using matrix operations, example solutions involving equivalence testing, vector addition, and scalar and matrix multiplication are available for download from the UPEP Curriculum Modules site [2] in the following systems: *MATLAB*, *Mathematica*, and C with MPI for high performance computing. (Blue Waters Student Intern Jesse A. Hanley implemented the latter.) Several datasets for use in projects also accompany the module.

2.2 High Performance Computing in Module

In line with the aims of UPEP, the module has an introductory section on "Population Matrices and High Performance Computing" that discusses the need for high performance computing (HPC) within the context of an ecological study of blue crabs. One such study has collected over a terabyte of data (10^{12} characters), and the researchers estimate that simulations would take about 5.7 years on a sequential computer. As the module points out, "With such massive amounts of data and such intensive computations, researchers must used high performance computing with multiple computer processors to store the data and large matrices and to perform the needed simulations in a reasonable amount of time" [6].

The section on "Population Matrices and High Performance Computing" can be covered for information only, as a starting point for class discussion, or as motivation for the students' own HPC project development. Moreover, students can develop sequential or high performance computing versions of the projects. For example, three projects use synthetic datasets for the activities of the population of Portland, Oregon, generated from real data by the Network Dynamics and Science Simulation Laboratory (NDSSL) at Virginia Technical University [7]. One of the module's projects requires high performance computing to process NDSSL's synthetic data involving 1,615,860 people having 8,922,359 activities.

2.3 Module Content and Applications

Matrices, vectors, and operations involving these data structures are essential to many network/graph-based computational science models. Thus, the module introduces the following important and foundational mathematical concepts: vectors, vector addition, multiplication of vectors by a scalar, dot product, matrices, scalar multiplication, matrix sums, matrix multiplication, square matrices, and the association of matrices and systems of equations.

These concepts are introduced in an environment of biological applications. As indicated above, the first section motivates the study of vectors and matrices and the need for high performance computing (HPC) with a discussion of a scientific study of blue crabs that includes high performance computing simulations.

The foundational material includes a discussion of vectors with such terms as "element," "size," and "index" and such concepts as equality and addition of vectors and multiplication by a scalar. Examples involve vectors of simulated changes in populations of competing white tip reef sharks and black tip sharks in an area.

Then, a section on "Dot Product" uses another biological example in estimating the number of eggs laid by Hawaiian green sea turtles in one year. As indicated in the module, scientists around the world have studied the magnificent green sea turtle and used mathematics and computer science to make predictions about their populations in efforts to keep these them from extinction. A figure features the Caribbean Conservation Corp John H. Phipps Biological Field Station Costa Rica for the study of the green sea turtle. The section's example begins: "We deal with a different type of multiplication in estimating the number of eggs laid by Hawaiian green sea turtles in one year. We can consider their life cycle to be in five stages with egg layers in two stages, novice breeders of age 25 years and mature breeders from age 26 through 50 years. On the average, a novice breeder lays 280 eggs in a year, and a mature breeder lays 70 eggs per year. We can combine these data in a vector $\mathbf{e} = (280, 70)$. Suppose also that there are 291 novice and 9483 mature breeders, which we store in the vector $\mathbf{b} = (291, 9483)$. To approximate the total green sea turtle egg production in a year, we multiply together corresponding terms and add the results, as follows:

$$\begin{aligned} \mathbf{e} \cdot \mathbf{b} &= (280, 70) \cdot (291, 9483) \\ &= 280 \cdot 291 + 70 \cdot 9483 \\ &= 81{,}480 + 663{,}810 \\ &= 745{,}290 \text{ eggs} \end{aligned}$$

This type of multiplication, the **dot product**, involves two vectors of the same size and results in a number, *not* another vector."

After a discussion of the option of writing a dot product with the first term as row vector and the second as a column vector, the section concludes with the following quick review question, whose answers are in a section at the end of the module:

"**Quick Review Question 4** The first stage in the life of the Hawaiian green sea turtle, consisting of eggs and hatchlings, occurs during the first year. Stage 2, juveniles, extends from year 1 to 16. Suppose 23% of the hatchlings survive and move to stage 2, while 67.9% of those in Stage 2 remain in that stage each year. In one year, suppose Stage 1 has 808,988 individuals, and Stage 2 has 715,774 (Green Sea Turtle).

- **a.** Give a vector, **p**, with real number elements representing the percentages.
- **b.** Give a vector, **s**, storing the individuals in Stages 1 and 2.
- **c.** Using variables **p** and **s**, not the data, give the vector operation to determine the number of individuals that will be in Stage 2 the following year.
- **d.** Calculate this value."

With the foundation of vectors and vector operations, the next three sections lead the student carefully through explanations of definitions involving matrices, scalar multiplication, matrix sums, and matrix multiplication using examples with populations of white tip reef sharks and black tip sharks and additional quick review questions.

A subsequent section on "Square Matrices" defines "square matrix" and "diagonal element." These terms are illustrated with hypothetical data of the distribution of ABO blood types (A, B, AB, O) for mothers and newborns (multiple births omitted) in a county over a year. In another example involving a square matrix, a table (Table 1 here) presents similarity measures (specifically, Euclidean distances) of the 18S rRNA sequences of pairs of animals, where smaller numbers indicate closer relationships. Using this example as motivation, the section defines "symmetric matrix".

Table 1. Similarity measures (specifically, Euclidean distances) of the 18S rRNA sequences of pairs of animals (Table 3 in Lockhart, 1994)

	Frog	**Bird**	**Human**	**Rabbit**
Frog	0	0.316	0.350	0.336
Bird	0.316	0	0.130	0.102
Human	0.350	0.130	0	0.028
Rabbit	0.336	0.102	0.028	0

For the concluding section on "Matrices and Systems of Equations," we return to the earlier example involving the dot product, where a Hawaiian green sea turtle novice breeder lays an average of 280 eggs per year, while a mature breeder only lays 70. The material continues, "Instead of specifying the number of turtles in each category, let n be the number of novice breeders and m the number of mature breeders with $\mathbf{b} = (n, m)$. In general, the average annual egg production, a, is computed as follows:

$$\mathbf{e} \cdot \mathbf{b} = (280, 70) \cdot (n, m)$$
$$= 280n + 70m = a$$

Thus, the dot product translates into one side of a linear equation." Moreover, the section indicates that a matrix-vector product involving the black-tip/white-tip shark application is equivalent to a system of linear equations.

2.4 Module Exercises and Answers

After the body of educational material, the module contains a section with 75 exercises. Instructions encourage students to check their work with a computational tool, while a subsequent section has answers to 15 exercise parts. Many exercises are routine practice problems with vectors and matrices. However, several "word problems" involve applications, such as real spectrophotometer readings to indicate the number of bacteria in a broth; development of an age-structured model for an animal; a threshold matrix; and a dither matrix for enhancement of a digital image, such as a medical image from a CT (computerized tomography) scan.

2.5 Module Projects

After the reinforcement of concepts in the exercises, seven large projects are available for students to complete as individuals or with a team. Instructions indicate to use a computational tool, optionally with high performance computing except as indicated.

Three projects consider matrices and vectors as part of network-based epidemiology simulations. The current module can serve as a basis for another Blue Waters module by the authors, "Getting the 'Edge' on the Next Flu Pandemic: We Should'a 'Node' Better," which develops this graph theory based model in detail [8]. In preparation for development of simulations in this subsequent module or as stand-alone applications, the three projects lead students through various aspects involving vectors and matrices, such as formation of vectors of IDs for people and locations and of a corresponding people-to-people connection matrix.

Four projects employ colon crypt data that were output from simulations by Ornella Cominetti and the authors. The simulations were performed with Chaste (Cancer, Heart and Soft Tissue Environment), "a general purpose simulation package aimed at multi-scale, computationally demanding problems arising in biology and physiology" that a team centered in the Computational Biology Group at Oxford University Computing Laboratory is developing [9]. Scientists believe that colorectal cancer originates in tiny crypts that descend from the colon's epithelium into the underlying connective tissue. Projects, which use data downloadable from the website [2], are variations of those employed in research at Oxford and include plotting the trajectory of a cell in the crypt; generating a stacked bar chart of the average number of cells in various categories by time; plotting the mean migration velocities of cells at different heights in the crypt; and plotting the mean spatial correlation of velocity, a metric of the amount of coordinated movement of the cells, along with standard error bars. The projects develop the background necessary for their completion, and instructions ask the students to discuss their results.

2.6 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 a parallel version of a program using C and MPI to support this and other modules. His program accompanying the module is available on the NCSI UPEP Curriculum Modules site [2]. Jesse's experiences as an intern and program developer enhanced his understanding of programming in general and HPC in particular.

3. TESTING AND EVALUATION

3.1 Class Testing in High Performance Computing

The first author taught Wofford College's High Performance Computing (HPC) course (COSC 365) in the spring of 2010. A mixture of Emphasis in Computational Science (ECS) students and computer science majors (five students: one biology/ECS, one chemistry/ECS, one computer science/chemistry/ECS, two computer science; sophomore to senior level) populated the class. All students had taken Data Structures with programming in Python and C++ and at least one other computer science or computational science course.

In preparation for discussion of HPC implementation of matrices and of applications involving matrices, the class was assigned reading of a preliminary version of the "Living Links" module, various graded and non-graded exercises from the module, and a quiz with questions taken from its Quick Review Questions. This background helped to prepare the students to develop several projects, including population dynamics model of skates, which

are similar to sharks; performance analyses of sequential and parallel programs that raise square matrices of various sizes to a range of powers, run on a local cluster, NCSA's Teragrid computer Abe (a Dell Intel 64 Linux Cluster), and NICS' Teragrid computer Kracken (a 99072-processor Cray XT5 computer) [10]; and a social network/individual-based epidemiology simulation using another Blue Waters curriculum module by the authors, "Getting the 'Edge' on the Next Flu Pandemic: We Should'a 'Node' Better" [8].

3.2 Evaluation in High Performance Computing

The students demonstrated their understanding of the module's material with their performance on homework and a major exam. Feedback, both formal and informal, from class members about the preliminary version of the model helped revisions and formulation of the completed version.

3.3 Class Testing in Linear Algebra

Linear Algebra at Wofford College is a sophomore-level class required of mathematics majors and minors and computer science majors. In spring 2011, after covering the basics of vector operations, the instructor, Dr. Joseph Spivey, required that the students read the module before discussing matrix multiplication and its applications in class. The professor used one 50-minute class for the module but mentioned that had he covered everything in the module, the material would have taken one-and-a-half to two full periods. After the class, he assigned a number of the exercises to be done by hand.

3.4 Evaluation in Linear Algebra

Immediately after using the material, students in Linear Algebra along with their professor completed a questionnaire about the module. The questionnaires for the Linear Algebra class 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 responses of the 20 students in the class, which were between 4.30 and 4.50, reflect favorably on the module (see Table 2). For similar questions, the professor gave ratings of all fives (5s).

Table 2. Means and standard deviations of $N = 20$ responses to rated questions, 1 (strongly disagree) to 5 (strongly agree)

Question	Mean	Standard Deviation
I understood the science applications in the module.	4.30	0.47
I understood the mathematics in the module.	4.50	0.69
The module was readable.	4.33	0.69
The Quick Review Questions helped me understand the material.	4.38	0.84
The exercises helped me understand the material.	4.34	0.68

Students and the faculty member were also asked to elaborate about the above scores. Dr. Spivey commented, "I was impressed with how much they understood even before I went over it in class. Some students said to me that they learn better through the use of applications and that the math was explained very well." Student comments reflected the same impressions: "I felt the module was very well put together in a way that was easy to follow with just enough breaks in text to keep track of equations but not enough to lose track of the subject." "I like the idea of using math as a tool to help predict how our actions may affect populations or show how past actions did." "The applications allowed for better understanding of the mathematics material itself."

Participants were also asked to indicate what they liked best about the module. The instructor said, "The Quick Review Questions were nice, and several students wrote about that in their journal entries." Some class members indicated on their questionnaires that they liked these best, and one student wrote, "The Quick Review Questions served both as a learning tool and a reference for the exercises as it gave me practice, and its solutions pages in the back showed enough for me to really grasp the processes of some problems…." Other students liked best the "readability" of the module, "the explanations," the arrangement of the material, and "the real-world applications." As one student wrote, "What I liked most about the module were the real life and science applications that involved the topics, many of which I am also studying in my science courses. These references kept the topics intriguing."

The questionnaire also asked what they found most difficult about the module and to give corrections and suggestions for improvement. For revision, Dr. Spivey suggested, "You may want to make it clear at the beginning that the answers to the Quick Review Questions are at the end." In response to this comment, we added such a paragraph statement before the first question about the location of the answers and how to use the Quick Review Questions and answers as a learning tool. As the instructor suggested, we also revised the phrasing of one of the exercises and added four matrix multiplication exercises. Two of the students requested a section for answers to selected exercises. In response to this suggestion, we added such a section with answers to fifteen (15) exercise parts.

In reply to the questionnaire's request for further comments, Dr Spivey summarized, "The students responded well to the material. They enjoyed the reading and could follow it easily. They also really enjoyed learning about the applications." Student comments were in a similar vein: "I think the module was wonderfully written and everything was explained in an easy-to-read way." "Their [The authors'] ability to relate vectors and matrices to sea creatures is astounding, and it gives the math an entirely new dimension. It shows the reader how widespread the

influence of mathematics is, while helping them learn new techniques and concepts." "I wish all math textbooks were easy to read and understand like this module!"

3.5 Workshop Testing

In February, 2011, Monash eScience and Grid Engineering Laboratory (MeSsAGE Lab), directed by Professor David Abramson, sponsored two computational science workshops at Monash University in Melbourne, Australia, for Ph. D. students. Dr. Bob Panoff, Executive Director of the Shodor Foundation [11], was the main leader of the first week-long workshop, "Introduction to Computational Thinking," which had an afternoon session conducted by the authors on "Quantitative Modelling Using MATLAB: Introduction." Later in the month, the authors lead the second week-long workshop, "Computational Workshop for the Life Sciences: Bringing Computation to Life," in which half the time involved modeling using MATLAB, including a similar introduction. Other topics in this half were user-defined functions, looping, decisions, model fitting, and a day on parallel programming. The format of this part of the workshop was presentations interspersed with frequent exercises for participant pairs to complete. The presentations and exercises included a number of examples, applications, and projects from the "Living Links" module. For example, the class developed versions of three projects from the module: visualizing the trajectory of a simulated cell in a colon crypt; plotting the mean migration velocities of simulated crypt cells; and modeling a network of individuals in a community with matrices and vectors and computing in parallel the distribution of the number of contacts such individuals have with other people. Wikis [12] and [13] give presentation files, exercises sets, and more details about the workshops.

3.6 Workshop Evaluation

Questionnaires for the workshops were more general than those for the linear algebra class. With a rating scale of 1 to 4 for Poor to Excellent, respectively, two questions for participants in the second workshop, "Computational Workshop for the Life Sciences: Bringing Computation to Life," seem most relevant to module evaluation: "The clarity of the information provided was:" and "The program materials were:" High mean scores and participants' comments about the workshop, particularly the applications, were gratifying (see Table 3).

Table 3. Means and standard deviations of $N = 18$ responses to rated questions, 1 (Poor), 2 (Adequate), 3 (Good), 4 (Excellent)

Question	Mean	Standard Deviation
The clarity of the information provided was:	3.56	0.62
The program materials were:	3.56	0.62

4. CONCLUSION

"Living Links: Applications of Matrix Operations to Population Studies" and its associated programs in MATLAB, Mathematica, and C/MPI are currently available on the UPEP Curriculum Modules website [2]. Class testing in High Performance Computing (HPC) of a preliminary version of the module helped in its development and showed the utility of the module in introducing matrices and some of their applications as a component of a HPC course. Class testing in Linear Algebra lead to refinement of the module and demonstrated its value in introducing matrix and vector operations with numerous applications to a mathematics class. Class testing the module as a base for workshop lectures, exercises, and projects illustrated its value as a resource for a faculty member. High questionnaire scores and enthusiastic comments from undergraduate level computer science, mathematics, and computational science students and graduate level science workshop participants verify the conclusion that "Living Links: Applications of Matrix Operations to Population Studies" can be an effective educational module in a variety of classes, levels, and settings.

5. ACKNOWLEDGEMENTS

We would like to acknowledge the generous help of many people and organizations. The National Computational Science Institute Undergraduate Petascale Education Program (UPEP), which is an NCSA Blue Waters project in collaboration with the National Computational Science Institute (NCSI) and national HPC programs, funded development of the module and supported UPEP intern Jesse Hanley, who implemented the module's HPC programs. NCSI under the direction of the Shodor Foundation with Executive Director Dr. Bob Panoff is hosting the module and associated materials on its website. The Monash e-Research Centre with Science Director Dr. David Abramson provided travel support for the authors during their stay in Australia and also organized and sponsored the workshops. Dr. Joe Spivey class tested the module in his Linear Algebra class. NCSA and NICS provided access to their Teragrid computers for the UPEP intern, the High Performance Computing class, and the first author. The authors worked with Ornella Cominetti at Oxford developing Chaste simulations that provide the foundations and data for four projects.

6. 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. 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.

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

[4] Monash eScience and Grid Engineering Laboratory (MeSsAGE Lab) at Monash University in Melbourne, Australia. https://messagelab.monash.edu.au/ Accessed 3/5/11.

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

[6] Taylor, Caz and Erin Grey. "Population Dynamics of Gulf Blue Crabs" 2010. Tulane University. http://leag.tulane.edu/PDFs/Grey-LEAG-4.28.10.pdf Accessed 10/13/10.

[7] NDSSL (Network Dynamics and Simulation Science Laboratory, Virginia Polytechnic Institute and State

University). 2009. "NDSSL Proto-Entities" http://ndssl.vbi.vt.edu/opendata/ Accessed 8/27/9.

[8] Shiflet, A. and Shiflet, G. 2010. "Getting the 'Edge' on the Next Flu Pandemic: We Should'a 'Node' Better." National Computational Science Institute Undergraduate Petascale Education Program (UPEP) Curriculum Modules, UPEP Curriculum Modules site. http://shodor.org/petascale/materials/UPModules/socialNetworks/ Accessed 5/21/11.

[9] Chaste, Cancer, Heart and Soft Tissue Environment. 2010. http://web.comlab.ox.ac.uk/chaste/ Accessed 10/14/10.

[10] Teragrid. 2010. https://www.teragrid.org/ Accessed 3/5/11.

[11] Shodor, a national resource for computational science education. 2011. http://www.shodor.org/ Accessed 3/5/11.

[12] Panoff, R., Shiflet, A. and Shiflet, G. "Introduction to Computational Thinking." 2011. https://messagelab.monash.edu.au/IntroductionToComputationalThinking/ Accessed 3/5/11.

[13] Shiflet, A. and Shiflet, G. "Computational Workshop for the Life Sciences: Bringing Computation to Life." 2011. https://messagelab.monash.edu.au/ComputationalThinkingForLifeSciences Accessed 3/5/11