DIMACS Series in Discrete Mathematics
and Theoretical Computer Science

Why BioMath? Why Now?

Fred S. Roberts

Abstract. The rationale for and the opportunities to bring the biology-mathematics interface into the high schools are explored through examples from the author's career. Among topics considered are (1) epidemiological modeling, (2) biology as information science, (3) physical mapping of DNA, (4) DNA and RNA chains and the "RNA detective game," (5) systems biology, (6) graph-theoretical models of the spread of disease, (7) measurement of cough severity and fatigue, (8) biosurveillance, and (9) location of bioterrorism sensors. With each, activities appropriate for the high school Biology and Mathematics classrooms, and often both in partnership, are described.

1. Introduction

In 2002, I was invited to join the Secretary of Health and Human Services' Smallpox Modeling Group. How was a mathematician selected to join this group? The answer has a lot to do with the theme of this book.

The story I will tell is organized around my own career. One of the goals of bringing the biology-mathematics interface into the high schools is to open up to students the possibilities of new types of careers and in particular careers that blend disciplines. In turn, the availability of such careers is connected to new, rapidly-increasing, interdisciplinary educational programs in biomath at the college and graduate school level. Having examples of areas in which new types of careers are possible can be very helpful. In the process of telling this story, I will give a variety of examples of activities that are appropriate for high schools. I have tried many of them already with high school teachers and/or students. The blending of disciplines provides a wonderful vehicle for highlighting each of them, and the examples I give can be used in either high school Biology or high school Mathematics classes, and often in both with teachers collaborating in new partnerships.

One of the reviewers of this article felt that it would be helpful for me to summarize my educational background and career path. By way of background: My Bachelors, Masters, and Ph.D. degrees are all in Mathematics, though already as an undergraduate I had interests in the connections between Mathematics and other disciplines. After getting my Ph.D., I did a postdoctoral fellowship in a Psychology Department, worked in the Math Dept. at a "think tank," did another postdoc in

The author thanks the National Science Foundation for its support under grants xxxx, yyyyy, zzzzz.

©2010 American Mathematical Society

a social sciences group, and then joined the Rutgers University Mathematics Department, where I have been on the faculty since 1972. I have been associated with the Center for Discrete Mathematics and Theoretical Computer Science (DIMACS) since 1988. DIMACS, based at Rutgers, was one of the original National Science Foundation "science and technology centers" and is a consortium of academic, industry, and government partners (http://dimacs.rutgers.edu). Since 1996, I have been Director of DIMACS, which runs a wide gamut of research and education programs with an emphasis on interdisciplinarity.

2. Epidemiological Modeling

My story starts with mathematical modeling of epidemics, though that is more near the end than the beginning. Epidemiological models of infectious diseases go back to Bernoulli's mathematical analysis of smallpox in 1760 (Bernoulli [1760]) and since then mathematical models have been developed for key pathogens such as influenza (see e.g., Lin, Andreasen, and Levin [1999]), malaria (see e.g., Anderson and May [1991]), gonorrhea (see e.g., Hethcote and Yorke [1984]), tuberculosis (see e.g., Blower, et al. [1995]), and HIV (see e.g., Perelson, et al. [1996]). Understanding infectious systems requires being able to reason about highly complex biological systems, with hundreds of demographic and epidemiological variables. Intuition alone is insufficient to fully understand the dynamics of such systems. Experimentation or field trials are often prohibitively expensive or unethical and do not always lead to fundamental understanding. Therefore, **mathematical modeling** becomes an important experimental and analytical tool. Mathematical models have become important tools in analyzing the spread and control of infectious diseases, especially when combined with powerful, modern computer methods for analyzing and/or simulating the models. Great concern about the deliberate introduction of diseases by bioterrorists has led to new challenges for mathematical modelers. Great concern about possibly devastating new diseases like avian influenza or H1N1 virus ("swine flu") has also led to new challenges for mathematical modelers. Since Bernoulli's pioneering work, mathematical modeling has provided insights into drug-resistance, rate of spread of infection, and effects of treatment and vaccination, and is being used today to help us deal with emerging disease threats such as SARS or pandemic flu.

The size and overwhelming complexity of modern epidemiological problems calls for new approaches and tools. As a result, in 2002, DIMACS launched a "Special Focus" on "Computational and Mathematical Epidemiology" that has paired mathematicians, computer scientists, and statisticians with epidemiologists, biologists, public health professionals, physicians, etc. (See http://dimacs.rutgers.edu/-SpecialYears/2002_Epid/.) The special focus has featured tutorials, workshops, and research working groups and has introduced numerous students and faculty to the subject of epidemiology and the possible approaches to epidemiological problems through mathematics, computer science, statistics, and other methods. This special focus has also led to a wide variety of related activities, including activities that are bringing materials at the interface between the biological and mathematical sciences into the high schools and training teachers to bring such materials into their classrooms.

While much of traditional mathematical epidemiology uses more advanced mathematical tools such as differential equations, a growing body of methods that

use tools accessible to high school students is being developed. These include graph-theoretical models of spread of disease (see section on that topic below) and combinatorial group testing to test for AIDS and other sexually-transmitted diseases (Du and Hwang [2000]). The biological background required to understand these topics is minimal and requires only some understanding of hosts, pathogens, incubation periods, etc. At DIMACS, we have had considerable experience with introducing these topics at the high school level and have found them to be very successful with both mathematics and biology students and teachers who do not have a major background in mathematics or one in biology. We started out in 2005 with a program called the DIMACS BioMath Connect Institute (BMCI) that was DIMACS' first attempt to bring together high school biology and mathematics teachers to explore topics in biomath together and design activities and prepare materials to bring back to their schools (http://dimacs.rutgers.edu/dci/2005/). BMCI continued in 2006 and then we developed the DIMACS BioMath Connection (BMC) that is concerned with developing modules at the interface between the biological and mathematical sciences that are usable in both Biology and Mathematics classes in the high schools. These modules are pilot- and field-tested, and we run programs to train teachers to implement them. To date, modules have been developed or are in the process of being developed in three areas: computational molecular biology, epidemiology, and ecology/population biology. (For more about BMC, see http://dimacs.rutgers.edu/BMC/.)

BMC has so far produced three modules in epidemiology that are aimed at both math and biology classes in the high schools. They have been pilot-tested and field-tested in the schools and are, at this writing, being revised for completion. The titles of these modules are: "Mathematical Modeling of Disease Outbreaks," "Imperfect Testing," and "Competition in Disease Evolution." The contents of these modules reflect a variety of important themes in epidemiology, from a mathematical point of view.

The module "Mathematical Modeling of Disease Outbreaks" introduces simple mathematical models that can answer some of the following questions: Will there be a flu outbreak this season? How many individuals will become infected? How long will it persist? Would vaccination prevent an epidemic? What other measures could be taken to prevent an epidemic? Two hypothetical infectious disease outbreak investigations serve as the motivating examples and storyline. Students receive an introduction to the basic characteristics of viruses and bacteria and the differences in treatment options for diseases they cause. Students then participate in a classroom simulation of the spread of disease to introduce the concepts behind the mathematical model before the details of the model are described. The model is constructed and the students have the opportunity to interactively see how changes in the parameters of the model change the pattern of the disease outbreak. For example, students can assume that a certain proportion of the population is vaccinated and compare the resulting outbreak to one where none of the population is vaccinated. At the end of the module, the students return to the outbreak investigations and are able to understand how epidemiologists might go about trying to answer the questions posed using mathematical models that incorporate some of what they know about the biology of the pathogen. The Mathematics topics are

arithmetic and state graphs, and the Biology topics are viruses, bacteria and the diseases they cause; vaccination, transmission routes and spread of disease.[1]

The module "Imperfect Testing" uses a case study approach to answer the following questions: What do the results of an imperfect medical test actually mean? How does one measure the effectiveness of a particular medical test or compare tests? How does this information affect public policy or personal decision making? The results of a mammogram, like those of many tests, are not always correct. A false positive test result may create unnecessary anxiety, while a false negative test result may result in a false sense of security. The students are presented with the case of an adult female who learns her mammography test is positive. They then discuss the possible implications or outcomes of a positive test result, given the properties of the test. These properties, which include sensitivity and specificity, help to determine the rates of incorrect test results and the predictive value of a test for a single individual. Next the woman has a genetic test where she learns she has the BRCA gene mutations associated with breast cancer. This leads to a dilemma for her daughter who must now decide if she will be tested for this BRCA allele. Since results from testing for this allele still do not completely determine whether or not she will develop breast cancer, the students now learn about the concept of relative risk. The Mathematics topics include probability, conditional probability, ratios, and graphing a rational function, and the Biology topics include genetic testing, genetic variation, ethical choices, decision making based on data interpretation, taking perspectives, and gold standards. This module can be used in classes in biology, anatomy, algebra I, and algebra II.

In the module "Competition in Disease Evolution," students learn to consider infectious diseases from the perspective of evolutionary biology on a basic level. They gain an understanding of how different methods of pathogen reproduction can greatly affect the evolutionary fitness of a disease. After learning to compute simple and conditional probabilities, students use this to calculate probable levels of exposure to a disease in a population, probabilities of infection given exposure, and expected population-level rates of disease incidence. The Mathematics topics include rounding real numbers to integers, and converting among fractions, decimal representations and percentages, and the Biology topics include disease transmission, evolutionary fitness, natural selection, and evolutionary competition. This module is appropriate for use in Pre-Algebra or Algebra 1 courses, or in any biology class (Introductory through AP) that covers concepts of evolution and/or reproductive fitness. After using this module in her school, Vicki Shirley, a teacher from Corinth, Mississippi, indicated that contributions to the disciplines of mathematics, biology and teaching do indeed overlap: "The chance to 'team-teach' such a real-world topic was beneficial to both our students and to us (the teachers). Debbie, my partner, knew the biology and I knew the math so when you put us together .... we were a dynamite duo! Having us both in the classroom together gave the students two experts and we were able to field any question they had about the module. Together we were able to truly help the students to understand that Math and Biology can be used together to solve real-world problems. (Before this

[1]This paragraph and the next two are taken from the BMC Project Annual Report to the National Science Foundation, September 2008 and, in turn, were taken from teacher materials for the modules.

module, the students didn't even know that Biology and Math were even remotely related!!!)"[2]

In describing the result of using these modules in her classroom, biology teacher Kathy Gabric from Hinsdale Central High School, Hinsdale, Illinois, told us[3]: "Bringing the modules into my classroom has really opened my eyes. Students that are not so keen on biology but love math are suddenly animated. Students that love biology but are not so great at math, begin to see that the two go hand in hand. As they say on TV 'numbers are everywhere'. The modules have taken the topics they cover to a whole new level. The visualization that the math allows results in fantastic discussions about what is really occurring."

In sum, mathematical epidemiology is a wonderful vehicle for bringing the interface between the biological and mathematical sciences into the schools.

3. Biology as an Information Science

I have long been interested in applications of mathematics. I became interested in mathematical problems in biology very early in my career, well before smallpox and mathematical epidemiology became a major interest of mine. As a graduate student at Stanford University in the 1960s, I worked on a problem posed by award-winning geneticist Seymour Benzer. The problem involved molecular biology.

Molecular biology is a prime example of the new biology. Many biological phenomena are coming to be viewed as involving the processing of information (Jackson [2005]). Computer and information science are playing an increasing role in modern computational molecular biology and were critical players in the scientific breakthrough of sequencing the human genome. We played a role in this at DIMACS through the DIMACS Special "Year" on Mathematical Support for Molecular Biology (1994-2000) (see http://dimacs.rutgers.edu/SpecialYears/1994_1995-index.html). During this six-year program, we invited biological and mathematical scientists to collaborate, held workshops and tutorials, and led many previously esoteric topics such as alignments, physical mapping, and phylogeny reconstruction to become central areas of research in computer science and their precise mathematical formulation to become a building block for development of progress in computational molecular biology. Through such partnerships between biological and mathematical scientists, a variety of topics in biology have come to be studied from an information science point of view; examples include gene finding and motif recognition, protein and RNA folding, protein structure prediction, and linkage analysis. Through partnerships between mathematical and biological scientists, major new areas of research have been developed, stimulated by the availability of massive amounts of new data, the integration of experimental methods with algorithmic methods, and the development of powerful new tools for modeling ever-more-complex biological systems. In recent years, the term "digital biology" (Morris, et al. [2005]) has come to be used to represent such trends in the biological sciences.

Just as ideas from computer science and mathematics have led to new biological ideas and research areas, biological ideas have inspired new concepts and methods in information science. Increasingly, for example, analogies with naturally occurring

[2]The feedback from the teacher came from the extensive evaluation we did of the field-testing of the module, and was reported by our evaluator, Professor Len Albright.

[3]Private communication via email, September 2008.

biological phenomena such as "swarming" have led to paradigms for new computer algorithms.

It is not surprising that many undergraduate and graduate students are studying topics at the intersection between the mathematical and biological sciences. A 2008 US-China Computer Science Leadership Summit, which I organized in Arlington, Virginia, brought this point home in a dramatic way. The participants were deans, directors, and department chairs of leading computer science departments in the US and China. A major portion of the program was devoted to the increasing interplay between computer science and biology and for the need to find new ways to train computer scientists to work in this area. There is similar interest in doing this for students in the biological sciences. Both themes have been emphasized in such conferences and reports as Hastings, et al., [2002], Hastings and Palmer [2003], Levin, et al., [1997], Palmer, et al., [2003]. A variety of sources have called for programs integrating mathematics and biology at the undergraduate level. The report BIO2010 (Board on Life Sciences [2003]) recommends that concepts, examples, and techniques from math and the physical and information sciences be included in biology courses and that biological concepts and examples be included in other science courses. The National Institute of General Medical Sciences at NIH has launched an initiative to incorporate more mathematics and physics in the biology curriculum. The Biomedical Information Science and Technology Initiative at NIH (see http://www.bisti.nih.gov/) has launched a variety of programs to support development of connections between the mathematical and biological sciences. The Mathematical Association of America report Math & Bio 2010 (Steen [2005]) describes efforts at the undergraduate level to reduce barriers to cross-disciplinary collaboration and activities.

While the interface between the biological and mathematical sciences at the undergraduate and graduate level has taken off, high schools have done little to introduce students to these interconnections. In April 2005, Midge Cozzens and I organized the first international conference on linking the biological and mathematical sciences in the high schools at DIMACS (http://dimacs.rutgers.edu/Workshops/-Biomath/). At that conference, a number of speakers emphasized the possibility that introducing high school students to topics in biomath will enhance the study of both disciplines. It is likely that students interested in mathematics will find biological applications as a motivation to study more mathematics, as they will see how math is useful. Similarly, students in biology who are exposed to modern math/computer science topics relevant to biology will come to understand the importance of understanding modern mathematics and computer science. In both cases, there will be an opportunity to introduce students to new career possibilities and to newly-developing opportunities for further study.

Bringing the interconnections between the mathematical and biological sciences into the high schools requires new curricular materials that teachers can use. Development of such materials is the principal goal of the DIMACS BMC program that was discussed above. Moreover, we need to train teachers to use these new materials, and this will involve exposing them to the interface between the disciplines. Teachers from different disciplines need to learn each others' language, open lines of communication, and develop new approaches for introducing cross-disciplinary topics. This was a major goal of the DIMACS BMCI program, and this goal continued in the DIMACS BMC program. Through BMCI and BMC, we

have experimented with methods to prepare teachers to use our interdisciplinary materials, and we have started to evaluate these methods, document them, and disseminate the documentation.

Molecular biology programs at BMC and BMCI have been built around such topics as global and local string alignment algorithms; the BLAST algorithm, including algorithms for protein sequences; FAST algorithms (FASTP, FASTA); PAM matrices; phylogenetic trees and tree parsimony; phylogenetic tree reconstruction and phylogenetic footprinting; and genome rearrangement. These topics use mathematical methods that are easily accessible to high school students and do not require sophisticated mathematical background. The basic mathematics required involves graph theory, counting principles, the basics of probability, and the notions of string, substring, and superstring from computer science. The biological background for the topics involves the basics of genomes, sequences, genes, introns and exons, the genetic code, transcription, and translation, and is readily explained.

Three BMC modules in computational molecular biology have been pilot- and field-tested through BMC. "Biomatrices (Evolution by Substitution)" deals primarily with evolutionary processes resulting in amino acid substitutions due to changes in DNA; it does not assume any prior knowledge about biology. The mathematical content includes single- and multi-stage probability events, disjoint and independent events, matrices, matrix multiplication, and powers of matrices. The module does not assume any prior knowledge about biology. Some very basic knowledge of chemistry would be useful. Mathematics prerequisites are decimal multiplication and percentage calculations. Other mathematics topics in this module are not assumed as prior knowledge, and are developed in such a way that students can learn them for the first time or refresh their prior knowledge of those topics. The module is written to provide students clear access to the problem under investigation. The lessons were developed with the assumption that students have no prior exposure to matrices, or Markov chains, and minimal understanding of probability. Practice problems are provided to reinforce the key concepts and to vary the relative amount of attention paid to the content in each lesson of the module. At the completion of the module, assessment questions are also provided to determine what the students have learned and are capable of doing.[4]

In the module "Genetic Inversions," students apply the basic concepts of DNA and evolution to a particular kind of genetic mutation. They play a game involving the rearranging of sequences by inverting subsequences. Next, they are challenged to develop and write an algorithm for carrying out inversions. Finally, an improved algorithm is introduced and analyzed. Students connect the algorithm with the concept of gene mutation, and with the evolutionary distances that separate different species of animals. There is no assumption of any background in either biology or mathematics so that the module should be appropriate in all high school classes.

The module "Spider Silk" asks students to apply knowledge of protein structure and function to pose and answer the fundamental question: What alignment of two sequences is biologically most meaningful? The module develops the basic mathematical principles that underlie computer programs used to align amino acids

[4]This paragraph and the next two are taken from the BMC Project Annual Report to the National Science Foundation, September 2008 and, in turn, were taken from teacher materials for the modules.

nearly instantaneously. After becoming familiar with spiders, their webs and their silks, students use graphs (networks), dynamic programming, and recursive thinking to model sequence alignments. Then students use the computer program Biology Student Workbench (BSW) (http://bsw-uiuc.net/) to align the amino acids of silks from different species of spiders and interpret their differences. "Spider Silk" can be used in its entirety in either a Biology class or a Math class. If less time is available, Days 1-3 can stand alone in a Math setting, or Days 1, 4, and 5 could be used alone in Biology. Ideally the module could be team-taught by a pair of math and biology teachers. This module is appropriate for use with Biology 1 students who have studied protein structure, synthesis, and function, and classification of organisms. It would be an appropriate capstone to the study of DNA/RNA and protein synthesis in biology. Mathematically, "Spider Silk" would be an appropriate unit of study in a discrete mathematics course, either as a self-contained unit or in conjunction with the study of graph theory or recursion. Alternately, it could be used in more traditional math courses to introduce discrete mathematics topics as enrichment or to examine recursion in an applied setting in preparation for the study of sequences.

All biology students should be exposed to topics such as these because it is difficult to appreciate modern biology without viewing it through the lens of the mathematics underlying information science. Mathematics students should also be exposed to such topics, to enhance their appreciation for the wide variety of uses of modern mathematics. While I was not exposed to these topics in high school, my first exposure to the biology-mathematics interface came as a sophomore at Dartmouth College, when I learned about predator-prey models. That exposure opened up many new horizons for me. I followed that in graduate school at Stanford University with a Ph.D. thesis that combined my growing interest in the social and behavioral sciences (also stemming from Dartmouth) with a biological motivation, and, as mentioned above, included work on a problem posed by geneticist Seymour Benzer.

4. Benzer's Problem

Benzer's problem was posed in the late 1950s and, briefly put, asked: How can you understand the "fine structure" inside the gene without being able to see inside? (This was in the days before gel electrophoresis and modern methods of sequencing genomes.) Classically, geneticists had treated the chromosome as a linear arrangement of genes. Benzer [1959, 1962] asked whether the same thing was true for the "fine structure" inside the gene. So, was the gene fundamentally linear (as in Figure 1)? Or did it have a circular structure (as in Figure 2)? A figure-eight structure (as in Figure 3)? At the time, we could not observe the fine structure directly. Benzer studied mutations. He assumed mutations involved "connected substructures" of the gene. By gathering mutation data, he was able to surmise whether or not two mutations overlapped. Consider the data in Table 1, where the i,j entry is 1 if mutations Si and Sj overlap and 0 otherwise. Figure 4 shows connected substructures along a linear curve that have the same overlaps as indicated by Table 1. The beginning and end of the ith substructure is indicated by points Si and Si. Asking students to do the same thing for other tabular data is an activity that is readily accessible at the high school level. Benzer's 1959 paper gave overlap data for a small portion of the genetic structure of a certain

virus, bacteriophage T4 (phage T4). The overlap data for 19 mutants of phage T4 is consistent with the hypothesis of linearity. This data and corresponding linear representation are shown in Roberts [1976]. Benzer's paper includes partial overlap data for 145 mutations of phage T4. This data is also consistent with the hypothesis of linearity. Having students look up Benzer's 19-mutation data and checking the linearity makes for a good exercise.

FIGURE 1. Linear Structure for Gene.

FIGURE 2. Circular structure for fine structure in the gene.

FIGURE 3. Figure-eight structure for fine structure inside the gene.

	S_1	S_2	S_3	S_4	S_5	S_6
S_1	1	1	0	0	0	0
S_2	1	1	1	1	0	0
S_3	0	1	1	1	0	0
S_4	0	1	1	1	1	0
S_5	0	0	0	1	1	1
S_6	0	0	0	0	1	1

TABLE 1. The i, j entry is 1 if the mutations S_i and S_j overlap, 0 otherwise. (The example is from Roberts [1976].)

Suppose we represent the tabular (matrix) information as a vertex-edge graph with vertices corresponding to the rows and columns of the matrix and the vertices corresponding to rows i and j joined by an edge if and only if the i, j entry of the matrix is 1. Then if we can find such corresponding substructures along a linear curve, we say that the graph is an **interval graph**. Interval graphs have been very important in genetics. Figure 5 gives an example. To show that it is an interval graph, we find intervals, one for each vertex, whose corresponding overlaps correspond exactly to the edges in the graph. Such intervals are shown in Figure 6. A good exercise is to show that the cycle of length 4 is not an interval graph. Nor is the graph of Figure 7. Today, there are efficient algorithms for recognizing interval graphs and they have a great many applications both in biology and in a variety

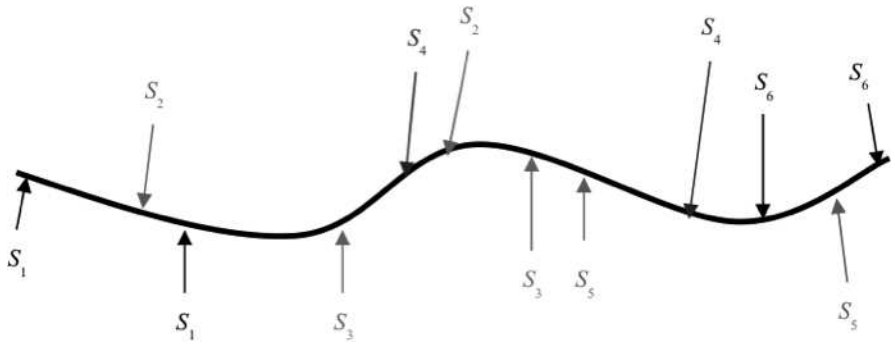

FIGURE 4. Connected substructures along a linear curve that have the same overlaps as indicated by Table 1. The beginning and end of the ith substructure is indicated by points S_i and S_i.

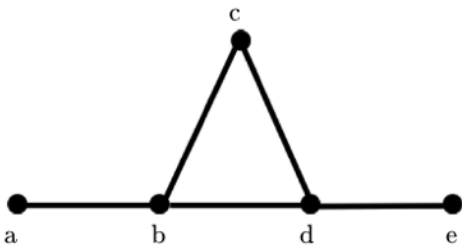

FIGURE 5. An interval graph.

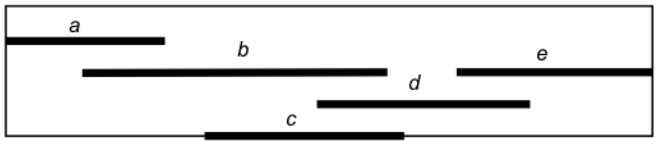

FIGURE 6. Intervals to show that graph of Figure 5 is an interval graph.

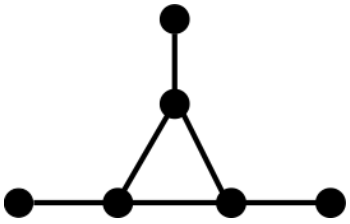

FIGURE 7. A graph that is not an interval graph

of other fields. For more on this topic, see for example Fishburn [1985], Golumbic [1980], or Roberts [1976, 1978].

Long after Benzer's problem was solved using other methods, interval graphs played a crucial role in **physical mapping of DNA** and more generally in the

mapping of the human genome. A physical map represents a piece of DNA, telling us the location of certain markers along the molecule, markers being precisely defined subsequences. In physical mapping, the first step is to make copies of the molecule we wish to map – the target molecule. Then we break each copy into disjoint fragments using restriction enzymes (more on fragments in the next section). We obtain overlap information about the fragments and then use the overlap information to obtain the mapping. One method of obtaining the overlap information is called hybridization. The fragments are replicated, giving us thousands of clones. The DNA fingerprinting is used to check if small subsequences called probes bind to fragments. The fingerprint of a clone is the subset of probes that bind to it. Two clones sharing part of their fingerprints are likely to have come from overlapping regions of the target DNA. From the overlap information, we create a **fragment overlap graph** whose vertices are fragments (clones) and with an edge joining two fragments (clones) if they overlap. If the clone overlap information is complete and correct, the fragment overlap graph is an interval graph. Then the corresponding "map" of intervals gives the relative order of fragments on the target DNA and this gives the beginning of a "physical map" of the DNA. For more details about this process, see for example Setubal and Meidanis [1997]; the preceding paragraph borrows heavily from their introduction to physical mapping.

5. DNA Chains and RNA Chains

So how did I get from Benzer's problem to modeling smallpox for the Secretary of Health and Human Services?[5] It has become increasingly clear, as I said above, that **biology has become an information science**. How so starts with deoxyribonucleic acid, DNA, the basic building block of inheritance and carrier of genetic information. DNA can be thought of as a chain consisting of bases. Each base is one of four possible chemicals: Thymine (T), Cytosine (C), Adenine (A), Guanine (G). Thus, using shorthand, the following are DNA chains:

GGATCCTGG, TTCGCAAAAAGAATC.

However, real DNA chains are long. That in Algae is 6.6×10^5 bases long; that in slime mold 5.4×10^7 bases long; that in a fruit fly 1.4×10^8 bases long; that in a chicken 1.2×10^9 bases long. DNA in humans is 3.3×10^9 bases long. The sequence of bases in DNA encodes certain genetic information. In particular, it determines long chains of amino acids known as proteins.

RNA is a "messenger molecule" whose links are defined from DNA. An RNA chain has at each link one of four bases, the same as those in DNA except that the base Uracil (U) replaces the base Thymine (T). Thus, GGAGUUCCAGU is an example of an RNA chain.

DNA and RNA provide prime examples of how fundamental mathematical methods of counting can be important in molecular biology and explained to high school students. Let us start by asking: How many possible DNA chains are there in humans? To answer this question, we apply the fundamental rule of combinatorics known as the Product Rule. Let us start with a simpler question: How many sequences of 0's and 1's are there of length 2? There are 2 ways to choose the first digit and, no matter how we choose the first digit, there are two ways to choose

[5]This section borrows heavily from Roberts and Tesman [2009]. All of the examples are taken from that book.

the second digit. Thus, there are $2 \times 2 = 2^2 = 4$ ways to choose the sequence. The four possible sequences are:

00, 01, 10, 11

How many sequences of 0's and 1's are there of length 3? By similar reasoning, we see that this is $2 \times 2 \times 2 = 2^3 = 8$.

This reasoning illustrates the **Product Rule**: If something can happen in n_1 ways and, no matter how the first thing happens, a second thing can happen in n_2 ways, then the two things together can happen in $n_1 \times n_2$ ways. More generally, if something can happen in n_1 ways and, no matter how the first thing happens, a second thing can happen in n_2 ways, and, no matter how the first two things happen, a third thing can happen in n_3 ways, ... then all the things together can happen in $n_1 \times n_2 \times n_3 \times \ldots$ ways.

So, how many possible DNA chains are there in humans? How many DNA chains are there with two bases? Using the product rule, we see that this is $4 \times 4 = 4^2 = 16$. There are 4 choices for the first base and, for each such choice, 4 choices for the second base. How many DNA chains are there with 3 bases? We get $4^3 = 64$. How many with n bases? We get 4^n. By this reasoning, the number of possible human DNA chains is $4 \wedge (3.3 \times 10^9)$, i.e., 4 to the 3.3×10^9 power. How big is this number? It is greater than $10 \wedge (1.98 \times 10^9)$ (1 followed by 198 million zeroes). A simple counting argument helps us to understand the remarkable diversity of life. Perhaps mathematical modeling will help us protect this rich array of life on our planet.

More sophisticated counting arguments can also help us to understand issues of molecular biology. RNA chains are very long. Early in the era of modern molecular biology, scientists asked if we can we discover what they look like without actually observing them. The trick they used was to split up long RNA chains into smaller ones, called **fragments**, using enzymes. The idea behind this leads to a mathematical challenge that I call the "**RNA Detective Game**," but which in the history of molecular biology was called the **fragmentation stratagem**. Some enzymes break up an RNA chain into fragments after each G link and others break up the chain after each C or U link. For example, consider the chain

Chain K: CCGGUCCGAAAG

Applying the G enzyme breaks the chain into the following fragments:

G fragments: CCG, G, UCCG, AAAG

We know that these are the fragments, but we do not know the order in which they appear. How many possible chains have these four fragments? Again using the product rule, we see that there are 4 choices for the first fragment, for each such choice 3 choices for second fragment, ... Thus, there are $4 \times 3 \times 2 \times 1 = 4! = 24$ possible chains. There is one chain corresponding to each permutation of these four fragments. One such chain different from the original is

UCCGGCCGAAAG

Suppose we instead apply the U,C enzyme to the chain K. We get the following fragments:

U,C fragments: C, C, GGU, C, C, GAAAG

How many chains are there with these fragments? Is $6! = 720$ the correct answer? Two of the permutations are the one that takes the fragments in the order given and the one that takes the second fragment first and the first second and all others

in this order. They give rise to the same chain. So 6! is wrong. What is the answer? What if the fragments were

C, C, C, C, C?

There are 5! permutations of these fragments, but only one RNA chain with these fragments:

CCCCC

To understand how to find the answer to this kind of counting problem, let us consider a classical combinatorics problem of putting n distinguishable balls into k distinguishable boxes. The number of ways to put n_1 balls into the first box, n_2 balls into the second box, ..., n_k balls into the k^{th} box is denoted by $C(n; n_1, n_2, \ldots, n_k)$, where $n = n_1 + n_2 + \ldots n_k$. It is well known that this **multinomial coefficient** is given by:

$$C(n; n_1, n_2, \ldots, n_k) = n!/n_1!n_2!...n_k!$$

Using this formula, we can calculate how many RNA chains of length 6 have 3 C's and 3 A's. Think of 2 boxes, a C box and an A box. How many ways are there to put 3 positions (balls) into the C box and 3 into the A box? (Similarly, if a family has 3 boys and 3 girls, how many different orders are there for the 6 children to be born?) The answer is $C(6; 3, 3) = 6!/3!3! = 20$. Some of these RNA chains are: CACACA, ACACAC, AAACCC. If a 6-link RNA chain is chosen at random, what is the probability of obtaining one with 3 C's and 3 A's? There are 4^6 possible RNA chains of length 6. Thus, the probability is given by

$$C(6; 3, 3)/4^6 = 20/4096 \approx 0.005.$$

The number of 10-link RNA chains consisting of 3 A's, 2 C's, 2 U's, and 3 G's is $C(10; 3, 2, 2, 3) = 25,200$. What if we know they end in AAG? Then, only the first 7 positions need to be filled, and 2 A's and one G are already used up. Hence, the answer is $C(7; 1, 2, 2, 2) = 630$. Notice how knowing the end of a chain can dramatically reduce the number of possible chains.

We can now return to the U,C fragments of the RNA chain K. We have already observed that the number of RNA chains with these fragments is not $6! = 720$. To calculate the answer, think of having 6 positions (there are 6 fragments) and assigning 4 positions to the C box, 1 to the GGU box, and 1 to the GAAAG box. Then, we see that the number of ways of doing this is given by

$$C(6; 4, 1, 1) = 6!/4!1!1! = 30$$

Actually, this computation is still a bit off, though not because the combinatorial argument is wrong. Notice that the fragment GAAAG does not end in U or C. Thus, we know it comes last. There are 5 remaining U,C fragments. The number of chains beginning with these 5 fragments is given by $C(5; 4, 1) = 5$. These chains begin as follows:

CCCCGGU, CCCGGUC, CCGGUCC, CGGUCCC, GGUCCCC

We get all chains with the given U,C fragments by adding GAAAG to the end of each of these:

CCCCGGUGAAAG, CCCGGUCGAAAG, CCGGUCCGAAAG,
CGGUCCCGAAAG, GGUCCCCGAAAG

Thus, there are 24 possible chains with the given G fragments and 5 with the possible U,C fragments.

However, we have not yet combined our knowledge of both G and U,C fragments. Which of the 5 chains above with the given U,C fragments have the right G fragments? CCCCGGUGAAAG does not: It has CCCCG as a G fragment. Checking the remaining 4 possible RNA chains with the given U,C fragments shows that only the third one has the given G fragments. Hence, we have recovered the initial chain. This is an example of recovery of an RNA chain given a **complete digest by enzymes**.

How remarkable is it that we could recover the initial RNA chain this way? How many RNA chains are there with the same bases as chain K? There are 12 bases: 4 C's, 4 G's, 3 A's, and 1 U. The number of chains with these bases is given by $C(12; 4, 4, 3, 1) = 138,600$. Thus, knowing the number of bases is not nearly as useful as knowing the fragments.

Consider another example. Suppose an unknown chain has the following fragments:

G fragments: UG, ACG, AC
U,C fragments: U, GAC, GAC

Does any fragment have to come last? The G fragment AC has to come last because it doesn't end in G. Thus, the other two G fragments come first in some order and there are only two possible RNA chains with these G fragments: UGACGAC, ACGUGAC. The latter has AC as a U,C fragment. So, the former is the correct chain.

Is it always possible to completely recover the original RNA chain given its G fragments and U,C fragments? The answer is no, i.e., that there are two distinct RNA chains with the same G and U,C fragments. Finding two such RNA chains makes for a good exercise, certainly appropriate for the high school classroom. It is one of many good exercises surrounding this "RNA Detective Game." For more on the RNA Detective Game, see Roberts and Tesman [2009].

The fragmentation stratagem we have described was used by R.W. Holley and his colleagues at Cornell in 1965 (Holley, et al. [1965]) to determine the first nucleic acid sequence. The method is not used anymore and was only used for a short time before other, more efficient methods were adopted. However, it has great historical significance and illustrates an important role for mathematical methods in biology. Currently, by use of radioactive marking and high-speed computer analysis, it is possible to sequence long RNA and DNA chains rather quickly. The mathematical power of the fragmentation stratagem, nevertheless, is a good illustration of the use of methods of discrete mathematics in modern molecular biology (and of the power of counting) at a level that is understandable in the high school classroom. Brother Patrick Carney of DePaul Catholic High School, Wayne, New Jersey reports (personal communication) having presented the RNA Detective Game to his students with considerable success.

6. Systems Biology

Simple counting arguments, as we have noted, can lead us to an appreciation of the remarkable diversity of life on earth. Modern mathematical methods allow us to deal with huge ecosystems and understand massive amounts of ecological data. Mathematical ecology and population biology has a long history. Yet, modern biology runs the gamut between understanding very huge systems

and very tiny systems. In 2003, the National Science Foundation and the National Institutes of Health asked me to organize a Workshop entitled Information Processing in the Biological Organism (A Systems Biology Approach). (See http://dimacs.rutgers.edu/Workshops/InfoProcess/ for information and a program of the Workshop.) The key thesis of the Workshop was that the potential for dramatic new biological knowledge arises from investigating the complex interactions of many different levels of biological information.

Such information starts from the small: DNA, RNA, and proteins. It extends to protein interactions and biomolecules, protein and gene networks, and from there to cells, organs, individuals, populations, and ecosystems. The Workshop investigated information processing in biological organisms from a systems point of view – the point of view in the modern area of research known as **systems biology**. The list of "parts" is a necessary but not sufficient condition for understanding biological function. Understanding how the parts work is also important. But it is not enough. We need to know how they work together. This is the systems approach. The Workshop was organized around four themes illustrative of the systems approach: Genetics to gene-product information flows; signal fusion within the cell; cell-to-cell communication; and information flow at the whole system level, including environmental interactions.

Some examples discussed at the Workshop will illustrate this approach. Princeton University professor Bonnie Bassler talked about her work on information processing between bacteria that helps squids maneuver in the dark. Bacteria process the information about the local density of other bacteria through a chemical language that allows them to "count" numbers of other bacteria and react as a population when this number has reached a "critical mass." They use this, for example, to produce luminescence. The process involved can be modeled by a mathematical model involving **quorum sensing**. Similar quorum sensing has been observed in over 70 species. For more on this topic, see http://www.hhmi.org/research/-investigators/basslerbio.html and http://www.molbio.princeton.edu/index.php ?option=content&task=view&id=27 (Quorum sensing is anticipated to be the topic of a future module in the BMC project.)

A second example involved a feedback system called the "P53-MDM2" loop, which is used in DNA damage repair. This was presented by Uri Alon, Weizmann Institute, and Galit Lahav, Harvard University. Alon and Lahav and their collaborators (Lahav, et al., [2004]) modeled the process by which the P53 - MDM2 feedback loop contributes to the regulation of DNA damage repair. This loop results in a warning signal when there is stress to the DNA and the system then decides whether to repair the damage or allow a cell to die – the death of a cell might protect the life of the organism.

A third example, based on the work of Raimund Winslow of Johns Hopkins University, was concerned with mathematical modeling of systems in the body such as the heart, specifically of phenomena arising in excitation/contraction coupling in the ventricle. Winslow's models (Hinch, et al. [2004]) study the behavior of calcium release channels, which are understandable using stochastic models based on notions of probability. The work has application to the connection between heart failure and sudden cardiac death.

These three examples involve more sophisticated mathematics than the counting ones, in fact dynamical systems and differential equations. However, at some

level, these too can be explained to high school students. Feedback loops can be represented as vertex-edge graphs. Quorum sensing involves simple concepts like majority rule. The stochastic nature of calcium release and other biological phenomena can be discussed at the level of simple probabilities. What is most important here is the wide variety of biological phenomena that can be discussed using mathematical language.

7. Graph-theoretical Models of the Spread of Infectious Disease

So, how did I get to model smallpox? In 2001, a group of us at DIMACS was discussing the news that "mad cow disease" had been identified in a cow in the U.S. One of my colleagues suggested that, since mathematicians were smart, maybe they could apply their methods to understand this disease, which threatened the food supply, the health of both cattle and humans, and the economy. Further discussions led to a plan for a DIMACS 5-year program on mathematical and computational epidemiology that would involve workshops, tutorials, and research groups on a wide variety of mathematical problems arising from trying to understand the spread of infectious disease. This was the DIMACS special focus on Computational and Mathematical Epidemiology mentioned earlier. Diseases to be studied included AIDS, malaria, influenza, etc. Workshops were planned on evolution of viruses, vaccination strategies, and methods of data mining for early detection of disease. We planned to initiate this program in Fall 2002. However, the September 11, 2001 World Trade Center terror attacks and subsequent anthrax attacks led us to rethink this. We were all set to go, and started earlier than planned, with an emphasis on bioterrorism motivated by the anthrax attacks. Since smallpox is one of the diseases that is considered a potential bioterrorist threat, we were led to discuss this disease. We organized a research group on mathematical methods for defense against bioterrorism, and it was through connections made in that group that I was asked to serve on a federal smallpox modeling group. (I also wrote an article on challenges for the mathematical sciences in defense against bioterrorism; see Roberts [2003].)

While much of classical mathematical epidemiology is based on methods of differential equations and dynamical systems, work in other areas of mathematics more accessible to the high school audience is also relevant and growing – as I noted above. Diseases are spread through social networks. "*Contact tracing*" (identifying the contacts people might have with those who are infected) is an important part of any strategy to combat outbreaks of infectious diseases, whether naturally occurring or resulting from bioterrorist attacks. A simple way to model social networks is to use vertex-edge graphs. The vertices represent people and an edge between two people indicates that they have some contact. Assume that vertices are in different states that may change over time. Let $s_i(t)$ give the state of vertex i at time t. In the simplest case, we might consider two states, 0 = susceptible, 1 = infected. (This is an **SI Model**.) We assume that times are discrete: $t = 0, 1, 2, \ldots$ More complicated models involve more states, e.g., susceptible, infected, exposed, recovered, etc. (SEI and SEIR models). The Kaplan-Craft-Wein model for the spread/control of smallpox, published in 2003, involved 16 different states. This included different stages of the disease, immune (through vaccination or recovery from the disease), traced but not vaccinated, etc. (See Kaplan, Craft, and Wein [2003].)

In such graph-theoretical models of the spread of disease, a vertex moves from state to state every discrete time. But what kinds of movements are allowed? Once you are infected, can you be cured? If you are cured, do you become immune or can you re-enter the infected state?

In a simple model with only two states 0 and 1, we set a "threshold" k and assume that a person in state 0 at time t moves to state 1 if at least k of their neighbors in the graph are in state 1: You become infected if sufficiently many of your neighbors are infected. It is assumed that once you are in state 1, you never go back to state 0. We call this process an **irreversible k-threshold process**. It is illustrated in Figure 8, where we take $k = 2$ and show the progression over time of the state of infection of vertices in a graph. Here, we represent the infected vertices (state 1 vertices) with solid circles like c and d, and the uninfected vertices (state 0 vertices) with hollow circles like a and b. Notice that at time 1, vertices a and b change from state 0 to state 1 since they each have two neighbors in state 1. Also, by time $t = 2$, the situation is fixed and never changes again.

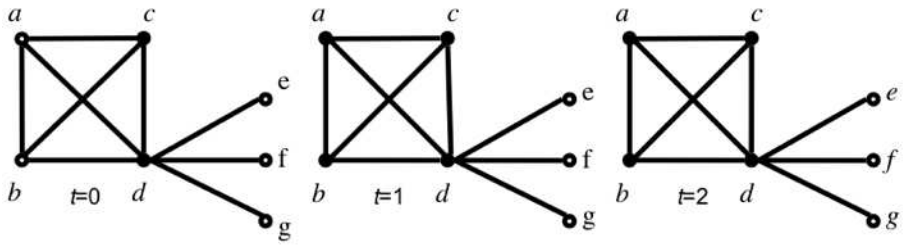

FIGURE 8. An irreversible k-threshold process with $k = 2$. Solid vertices are infected (state 1), hollow ones uninfected (state 0).

There are various complications we could add to this model, and I like to let my students suggest them. For instance, we could take $k = 1$, but only allow a vertex to get infected with a certain probability if it has an infected neighbor. We could add the condition that you are automatically cured after you are in the infected state for d time periods. We could give a public health authority the ability to "*vaccinate*" a certain number of vertices, making them immune from infection.

Mathematical models are very helpful in comparing alternative vaccination strategies. The problem is especially interesting if we think of protecting against deliberate infection by a bioterrorist. If you didn't know whom a bioterrorist might infect, what people would you vaccinate to be sure that a disease doesn't spread very much? In terms of the graph, we will think of vaccinated vertices as staying at state 0 regardless of the state of their neighbors. The question of whom to vaccinate makes a good exercise for students. Consider the graph of Figure 9, a "5-cycle." One strategy is **"mass vaccination"**: Make everyone 0 and immune in the initial state. In the 5-cycle, mass vaccination means vaccinate all 5 vertices. This obviously works to protect the population in the sense that no one can be infected. However, in practice, vaccination is only effective with a certain probability, so results could be different. Also, vaccines have side effects, so some people could get sick or even die if we vaccinate. If vaccine has no cost and is unlimited and has no side effects, of course we use mass vaccination.

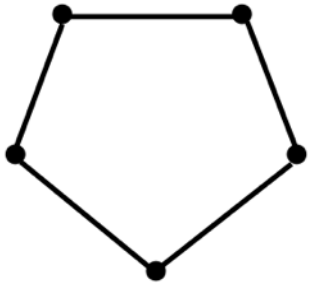

FIGURE 9. A 5-cycle.

What if vaccine is in limited supply? Suppose we only have enough vaccine to vaccinate 2 vertices. Suppose we can only vaccinate at time 0 and no new vaccine will become available after that time. Suppose an adversary has two doses of a "pathogen" that can be used to infect two vertices at time 0. Consider an irreversible 2-threshold process. There are, up to symmetry, two different vaccination strategies: Vaccinate two neighboring vertices and vaccinate two non-neighboring vertices. These are shown in Figures 10 and 11, with a V indicating vaccinated vertices. If you assume your adversary does not try to infect vaccinated vertices, the adversary has in each case, up to symmetry, two responses: infect neighboring vertices or non-neighboring vertices. The "alternation" between your choice of a defensive strategy and your adversary's choice of an offensive strategy suggests we consider the problem from the point of view of game theory. The Food and Drug Administration is studying the use of game-theoretic models in the defense against bioterrorism. In the graph of Figure 10, in an irreversible 2-threshold process, if the adversary infects two neighboring non-vaccinated vertices, no one else gets infected. However, if the adversary infects two non-neighboring non-vaccinated vertices, three vertices end up being infected. In the graph of Figure 11, in an irreversible 2-threshold process, no matter what the adversary does by way of infecting two non-vaccinated vertices at time 0, no one else gets infected. Thus, the vaccination strategy represented by Figure 11 is better in the sense that in the worst case fewer people end up getting infected. Your students could "play" with this type of problem for larger graphs, make assumptions that allow one to vaccinate more often or the adversary to infect people more often, etc.

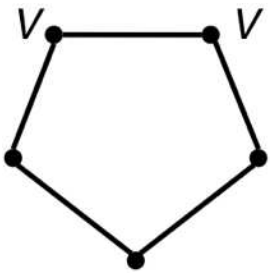

FIGURE 10. First vaccination strategy. V indicates vaccinated vertices.

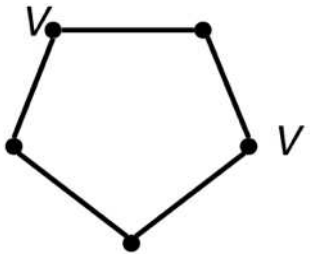

FIGURE 11. Second vaccination strategy. V indicates vaccinated vertices.

Suppose an adversary is out to infect as many people as possible. Given a graph, what subsets S of the vertices should he or she plant a disease with so that ultimately the maximum number of people will get it? This problem has an economic interpretation: What set of people do we place a new product with to guarantee "saturation" of the product in the population? As a defender against bioterrorism, your job can be defined as follows: Given a graph, what subsets S of the vertices should we vaccinate to guarantee that as few people as possible will be infected? A more extreme version of the attacker's problem is: Can we find a set of vertices to infect that will guarantee that ultimately everyone is infected? Mathematically, we speak of an **irreversible k-conversion set**: A subset S of the vertices that can force an irreversible k-threshold process to the situation where every state $s_i(t) = 1$. Note that if we can change back from 1 to 0 at least after awhile, we can also consider the Defender's Problem: Can we guarantee that ultimately no one is infected, i.e., all $s_i(t) = 0$? A variant of the Defender's Problem asks us to design a graph (with a given number of vertices and/or edges) that minimizes the number of vertices an opponent can ultimately force into the state 0.

To illustrate these ideas, consider the graph of Figure 12. It is easy to see that an irreversible 2-conversion set consists of the vertices x_1 and x_3. For if we infect them at time 0, then at time 1, x_2 gets infected, then at time 2, x_4 and x_5 get infected, and at time 3, x_6 gets infected.

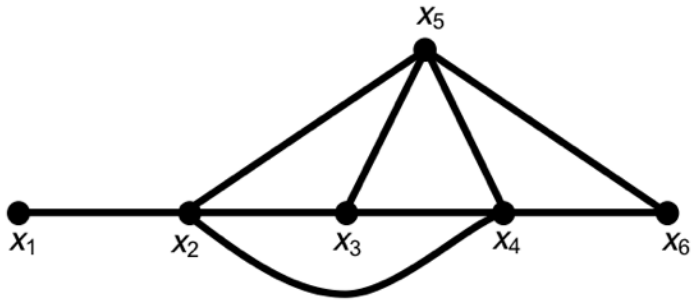

FIGURE 12. The set $\{x_1, x_3\}$ is an irreversible 2-conversion set.

One can prove a variety of theorems about irreversible k-conversion sets. See Dreyer and Roberts [2009] for details. Here, we simply note some simple examples that students might be challenged to investigate. The size of the smallest irreversible 2- conversion set in a cycle with n vertices is ceiling$[n/2]$, i.e., the least

integer greater than or equal to $n/2$. This is easy to verify, starting with examples. Finding irreversible k-conversion sets in trees and in grids also makes for interesting exercises illustrative of the concepts. Remarkably, even for grids in the plane, we don't know the size of the smallest irreversible k-conversion sets even for small values of k.

A variant on the vaccination problems we have discussed allows a defender to vaccinate v people **per time period**, while an attacker can only infect people at the beginning. What vaccination strategy minimizes the number of people infected? In the literature, this is sometimes called the **firefighter problem** and we think of a forest with trees, with a blaze spreading from trees to neighbor trees unless there is a firefighter placed at those trees. Consider an irreversible k-threshold process with $k = 1$, so a tree catches fire if any of its neighboring trees are on fire and it does not have a firefighter protecting it. This is a place to observe an important property of mathematical models: Once we translate a problem into mathematical language, the analysis can apply under multiple interpretations, whether it is firefighting or infection control. A variation on this problem in the infection context has the vaccinator and infector alternate turns, having v vaccinations per period and p doses of pathogen per period. What is a good strategy for the vaccinator?

To illustrate the process with $v = 3$ and $i = 0$ on a grid, consider the 8×12 rectangular grid, where point (i, j) means the point in row i and column j. Suppose the process starts with vertex $(4, 7)$ infected. One possible defensive strategy then is for the vaccinator to put firefighters at ("vaccinate") vertices $(3, 6), (3, 8)$, and $(5, 7)$. If this is done, then the fire spreads to vertices $(3, 7), (4, 6)$, and $(4, 8)$, the three "unprotected" neighbors of $(4, 7)$. Suppose the vaccinator then puts firefighters at $(2, 7), (4, 5)$, and $(5, 6)$. In turn, the fire spreads to $(4, 9)$ and $(5, 8)$. Suppose now the vaccinator puts firefighters at $(3, 9)$, $(4, 10)$, and $(5, 9)$. Now the fire can only spread to $(6, 8)$. In the final step, it is completely surrounded by putting firefighters at $(6, 7)$, $(6, 9)$, and $(7, 8)$. There are many issues to be studied in the firefighter/disease spread problem. In the language of firefighting, we might ask: Can the fire be contained? How many time steps are required before the fire is contained? How many firefighters per time step are necessary? What fraction of all vertices will be saved (burnt)? Does where the fire breaks out matter? What about a fire starting at more than one vertex? For some references on the firefighter problem, see Dreyer and Roberts [2009] and, for example, Hartnell and Li [2000], MacGillivray and Wang [2003], and Wang and Moeller [2002]. In the above example, 12 firefighters (doses of vaccine) are required to contain the fire (epidemic) and in the end, out of 96 vertices, seven are burnt and the rest are saved. Your students might want to see if they can do better than this either in terms of number of firefighters needed or in terms of more vertices saved.

One example of a problem that should be readily accessible to high school students involves fighting diseases (firefighting) on graphs that are given as trees that are rooted and that we can navigate from top to bottom. For each vertex u, define the weight$(u) = 1 +$ number descendants of u, where a **descendant** is a vertex reachable from u by a path heading downwards in the tree. Assume that the number of doses of vaccine (number of firefighters) per time period is given by $v = 1$ and the attacker infects the root and then cannot infect anyone after that. One algorithm for distributing doses of vaccine is the greedy algorithm: At each time step, place a firefighter (vaccinate) a vertex u that has not been saved such

that weight(u) is maximized. Consider the tree of Figure 13. A disease starts at the root. In the following process, the vertices that end up being infected are shown in the figure as black, while those that end up being noninfected are white. The vertices one level below the root are called level 1 vertices, and similarly for level 2, level 3, etc. The leftmost vertex at level 1 has weight 12 and that is highest, so it is vaccinated. At time $t = 1$, all other vertices at level 1 are infected but vertices descended from the vaccinated vertex can never get infected. Next, we consider vertices at level 2 that are not descendants of a vaccinated vertex. Two of them have the maximum weight 6. We choose at random in case of ties, say choosing the weight 6 vertex to the right to vaccinate, thus also saving all its descendants. The remaining level 2 vertices now get infected at time $t = 2$. At $t = 3$ we vaccinate one of the vertices at level 3 that is not a descendant of a vaccinated vertex and that has weight 3 and at time $t = 4$, we vaccinate one of the vertices at level 4 that is not a descendant of a vaccinated vertex. In the end, if this greedy procedure is used, 26 vertices are infected and 22 are not.

The greedy algorithm does not always lead to the result with the largest number of uninfected people. For example, consider the tree of Figure 14. If we use the greedy algorithm, we vaccinate the right-hand vertex at level 1 and end up with 7 uninfected people. However, if we vaccinate the left-hand vertex at level 1, we can end up with 9 uninfected people. Hartnell and Li [2000] showed that for any tree with one infection starting at the root and one dose of vaccine to be deployed per time step, the greedy algorithm always saves more than half of the vertices that any algorithm saves. Is this an acceptable solution? If we have a rapidly escalating epidemic, finding a speedy solution that is pretty good through an efficient algorithm might be preferable to finding an optimal solution if it takes so long to find the latter that it cannot be implemented in time.

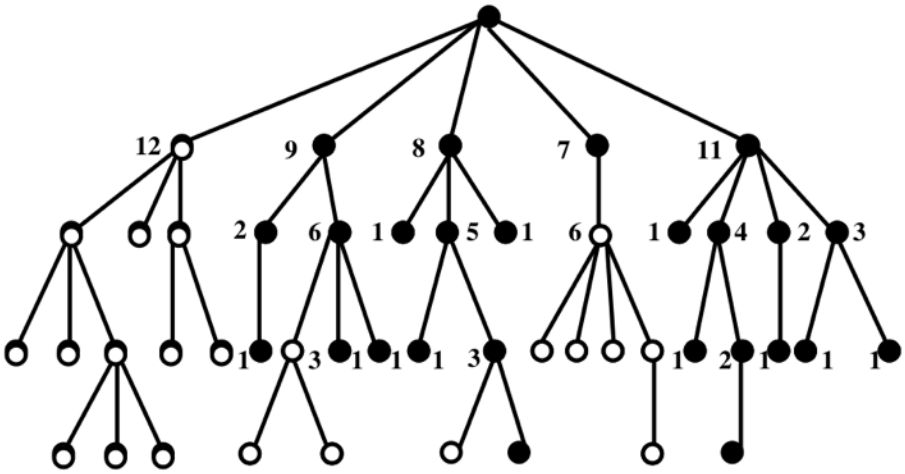

FIGURE 13. Tree for which the greedy algorithm leads to 26 infected and 22 uninfected vertices. The infected vertices are black, the uninfected ones white.

These graph-theoretical models of spread of disease have engaged me and my graduate students. At the same time, I have talked about them to high school

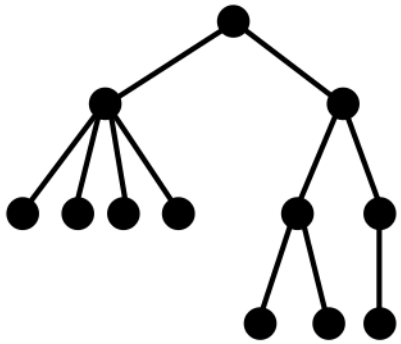

FIGURE 14. Tree for which the greedy algorithm does not lead to the smallest number of uninfected people.

audiences. One group of high school students at the Charter School of Wilmington developed smallpox models that led to their winning first place at the New Castle Science Fair and invitation to the International Science and Engineering Fair in Cleveland, Ohio. They were later invited to present their work to a group of researchers at a workshop on epidemiological modeling at DIMACS.

8. The African BioMath Initiative and Measurement Problems (Cough Severity, Fatigue)

Over the last two decades, the African continent has experienced devastating loss of life with catastrophic consequences, owing to the spread of deadly diseases such as HIV/AIDS, tuberculosis, malaria, and influenza. Diseases of Africa provide new and complex challenges for mathematical modeling. Because of modern transportation systems, no one in the world is safe from diseases originating elsewhere. Major new health threats such as H1N1 virus ("swine flu") or avian influenza present especially complex challenges to modelers in the context of developing countries. Such challenges were explored in several workshops during the DIMACS Special Focus on Computational and Mathematical Epidemiology and then led DIMACS to a major new African Biomathematics initiative (see http://dimacs.rutgers.edu/US-AfricanInitiative/). We have already run workshops and student short courses for U.S. and African participants on mathematical modeling of infectious diseases of Africa. The goals of our African initiative include studying challenges for mathematical models arising from the diseases of Africa; understanding the special challenges from diseases in resource-poor countries; bringing together U.S. and African researchers and students to collaborate in solving these problems; and laying the groundwork for future collaborations to address problems of public health and disease in Africa. The long-run future of collaborations on these problems rests in getting young people interested in them. This starts at the precollege level.

Our African Initiative has already led to some themes that should be of special interest to high school students. In several meetings/short courses in South Africa, we have worked on mathematical problems arising from diseases that inflict a significant burden on Africa. HIV/AIDS is a major case in point. Here, mathematics

has been important in evaluation of alternative preventive and therapeutic strategies; allocation of anti-retroviral drugs; evolution and transmission of drug-resistant strains of antibiotics; and the interaction of HIV/AIDS with other infections such as TB and malaria. (Such "coinfection" is a major theme in modern mathematical epidemiology.) Malaria is another interesting case in point. For malaria, mathematics has helped develop new methods of control (e.g., insecticide-treated cattle). It has also been used to study the effect of global warming on mosquito populations – a topic of central importance in a new DIMACS initiative on climate and disease. Children are often exposed to news about climate change and other similar problems of society. Many of these problems are fundamentally interdisciplinary and it is important for them to understand as early as possible how different disciplines interface.

Diseases of animals are of special significance in the developing world. In Africa, some key examples of such diseases are bovine tuberculosis (in domestic and wild populations), avian influenza, and trypanosomiasis. Since children are especially interested in animals, discussion of animal diseases has the strong potential of engaging them. Our workshops and short courses in Africa have included mathematical models of animal diseases. The U.S. Department of Homeland Security has established a university "center of excellence" FAZD (the National Center for Foreign Animal and Zoonotic Disease Defense), based at Texas A&M University, to study the spread of disease among animals. FAZD (see http://fazd.tamu.edu/) has an extensive educational program, emphasizing the importance of educating the next generation of homeland security workers in the area of disease.

While we hear about human diseases a great deal, and from time to time are exposed to diseases of animals, few of our young people understand that plants are also subject to disease. Diseases of plants are a major threat to the food supply not only in Africa but everywhere in the world. The Department of Homeland Security has also established a university center of excellence NCFPD (National Center for Food Protection and Defense), based at the University of Minnesota, that deals with protection of the food supply. A number of mathematical questions amenable to presentation at the high school level include the dose-response models needed to understand the impact on human health of certain concentrations of agents in food, and NCFPD, like FAZD, has an extensive educational program. (See http://www.ncfpd.umn.edu/ for more about NCFPD.)

DIMACS is running some new programs in its African Initiative. Each presents possible problems of interest at the high school level. A DIMACS workshop and shortcourse on conservation biology was held in South Africa in 2010. They dealt with the mathematics of ecological reserves. A key idea here that could lead to interesting discussion at the high school level is to define a notion of biological diversity that applies to a reserve with a variety of species and that provides a reasonable metric for the health of the reserve. Is biodiversity the inclusion of a lot of species? A reasonable distribution of individuals in each species? A distribution of individuals of varying ages? (See for example Sarkar [2002].) Another DIMACS workshop, to be held in Madagascar, will deal with genetics and disease control. In terms of food supply, this will deal with the safety of genetically altered crops. In terms of diseases, it will consider for example control of malaria by genetically modifying mosquitoes. One idea here is to sterilize male mosquitoes. What percentage of males do you need to sterilize to cut mosquito population significantly? What

if sterilized males are more attractive to females for mating – say 25% more? Or what if they are less attractive? How would this change the conclusions? (For more on this topic, see for example Adam [2005].) Still a third workshop and shortcourse was held in Uganda in 2009. It explored the relationship between economics and epidemiology. If people won't comply with a quarantine order, how much should we pay them to do so? What incentives will encourage more people to get tested for HIV? (For more on economic epidemiology, see e.g., Klein, et al. [2007]). We have followed up this workshop with an emphasis on "behavioral epidemiology" in the DIMACS Special Focus on Computational and Mathematical Biology, exploring ways to bring into epidemiological models the difference between individual responses to disease events, caused by their own personal economic considerations or their own priorities, concerns, and attitudes toward risk. These issues of economic epidemiology, genetics and disease control, and conservation biology can all be formulated in relatively simply mathematical models that should be of interest both in Biology and Mathematics classes at the high school level.

One of the major complications of HIV is the susceptibility of patients to other diseases, in particular tuberculosis. One activity that can be of interest even at very elementary precollege levels is to discuss how to measure the severity of the cough associated with TB. Another is to measure the fatigue associated with any disease such as HIV. Typically, we measure severity of a patient's cough on a 5-point scale: 5 = extremely severe, 4 = very severe, 3 = severe, 2 = slightly severe, 1 = no cough. To test a particular cough-suppressant, we might ask if the average cough severity in a group of patients treated with the suppressant is lower than the average in a "control" group. To make this precise, we let $a_1, a_2, \ldots, a_n$ be the patients in the first group and $b_1, b_2, \ldots, b_m$ be the patients in the second group. Let $f(x)$ be the severity of patient x's cough. Then we would like to see if

$$\frac{1}{n}\sum_{i=1}^{n} f(a_i) < \frac{1}{m}\sum_{i=1}^{m} f(b_i). \tag{1}$$

We are comparing **arithmetic means**. Note that there can be a different number of patients in each group, which is why we use n for one and m for the other. For instance, if there are three patients in the test group and five in the control group, those in the test group have cough severities 4, 3, and 1 while those in the control group have cough severities 5, 5, 2, 1, and 1, then the average cough severity in the test group is 2.67 while that in the control group is 2.8 and we conclude that, indeed, the average cough severity is less in the first group. On the other hand, if we were comparing the median cough severity, then the test group has median 3 and the control group median 2, and the conclusion about average cough severity is no longer true. We have to decide which of these two ways of averaging scores is the more appropriate one, which may depend on the application. (There are more complications. In the theory of measurement (Roberts [1979, 1994]), one distinguishes the types of scales used. In this case, it is possible to argue that the 5-point scale is "ordinal". In the case of ordinal scales, there are arguments that comparison of medians is "meaningful" in a precise sense whereas comparisons of arithmetic means is not.)

Judgments of cough severity are subjective. Suppose we ask a number of health care professionals to make a judgment of the severity of patients' coughs. We want to compare the average cough rating of patient a to the average cough rating of

patient b. Let $f_i(x)$ be the cough severity rating of patient x by health care worker i. Now, instead of Equation (1), we have

$$\frac{1}{n}\sum_{i=1}^{n} f_i(a) < \frac{1}{n}\sum_{i=1}^{n} f_i(b). \tag{2}$$

Note the subtle differences between Equations (1) and (2). In (2), we have the same n on both sides since there are the same number of raters in each case. Also, the subscript has moved to the f. We can make the same kind of analysis as before. Here, the exercise is to ask the students to produce the equation and to explain the differences.

If instead of cough severity, we talk about weight loss, then we could measure the weight in kilograms or in pounds. If we do that, then it is a good exercise to have the students show that if (1) is true in kilograms, it must also be true in pounds, and vice versa. It is also true that if (2) is true with all raters using kilograms, then it is also true if all raters use pounds, and vice versa. However, it is also an interesting exercise to ask students to give a numerical example to show that if raters can choose their scales, then (2) can be true with some raters using pounds and some using kilograms but false with some other combinations of who uses pounds and who uses kilograms. For example, you could ask the students to show via numerical example that if $n = 2$ and rater 1 uses pounds while rater 2 uses kilograms, then statement (2) could be true, whereas if rater 1 switches to kilograms and rater 2 switches to pounds, then (2) could fail. Surprisingly, the latter is not the case for geometric means and it is a good exercise to prove this. Here, we are talking about the statement

$$\sqrt[n]{\prod_{i=1}^{n} f_i(a)} < \sqrt[n]{\prod_{i=1}^{n} f_i(b)} \tag{3}$$

Similar examples can be given using the scale measuring fatigue. In serious diseases, one widely used scale is the Piper scale of fatigue. It asks questions like:

- On a scale of 1 to 10, to what degree is the fatigue you are feeling now interfering with your ability to complete your work or school activities? (1 = none, 10 = a great deal)
- On a scale of 1 to 10, how would you describe the degree of intensity or severity of the fatigue which you are experiencing now? (1 = mild, 10 = severe)

For more on averaging judgments of cough severity and of fatigue, see Roberts [2010].

9. Biosurveillance

As diseases spread rapidly from country to country or within a large country, and, for example, travelers from Africa can bring highly infectious diseases such as Ebola across the miles to the U.S. in a matter of hours, it becomes incumbent upon us to develop new ways of identifying the outbreak of diseases. Early on in my exposure to epidemiology, I learned about the importance of "biosurveillance." Biosurveillance is aimed at giving us early warning of the outbreak of a disease – whether naturally occurring or caused by a bioterrorist. This involves modern

data-gathering methods, which bring with them new challenges for mathematicians. There are some obvious types of data to gather in biosurveillance, including reports on number of people diagnosed with a given disease. However, before diagnosis, we would like to get early warning that something is happening. For this purpose, we want to look at a variety of data and see if, using various pieces of data, we can see patterns that suggest that there is a "disease event" taking place. Some of the unusual sources of data being considered today are: managed care patient encounter data; pre-diagnostic/chief complaint (emergency department data); over-the-counter sales transactions at drug stores or grocery stores; 911-emergency calls; ambulance dispatch data; absenteeism data; emergency department discharge summaries; prescription/pharmaceuticals; "hits" on certain medical information websites; and "adverse event reports" about responses to diseases and vaccines. For example, the New York City Department of Health (see e.g., Mostashari [2002]) looks for absenteeism among subway workers, who presumably would be exposed to a new disease in the city through subway riders. But it might combine this with other data. For instance, suppose there is an unusual amount of subway worker absenteeism in workers whose trains pass through a part of the borough of Brooklyn and there is also an increase in sales of Tylenol in the same neighborhood and an increase in "hits" on websites that deal with "achy legs." This "syndrome" of data might give warning that a certain event is taking place. Today, we sometimes talk about the special type of biosurveillance called "syndromic surveillance" (see http://www.cdc.gov/ncphi/disss/nndss/syndromic.htm) that includes combinations of symptoms and other observed phenomena. New methods of syndromic surveillance are being developed. They include spatial-temporal "scan statistics;" statistical process control (SPC); Bayesian applications; "market-basket" association analysis; text mining; rule-based surveillance; and change-point techniques. It should be easy to devise exciting classroom activities that include syndromic surveillance where we give each student some symptoms and do some "monitoring" to discover syndromes.

At DIMACS, our work on syndromic surveillance was carried out in a "working group" on disease detection, which led to collaboration with the CDC (U.S. Centers for Disease Control and Prevention). CDC has recently launched a new program on mathematical modeling of disease.

There are many sources of data that students can access and examine in classroom activities involving syndromic surveillance. Some examples are the following[6]:

- Morbidity and Mortality Weekly Report: http://www.cdc.gov/mmwr/
- SEER Cancer Registry: http://seer.cancer.gov/
- US Vital Statistics: http://wonder.cdc.gov/welcome.html

10. Bioterrorism Sensor Location

Early warning is critical in public health and that is a major reason for our emphasis on new tools for biosurveillance. Biosurveillance is important for all kinds of diseases, both naturally-occurring ones, ones that come around periodically like influenza, and those that are deliberately introduced by "bioterrorists." Smallpox is one of the diseases that the government fears could be reintroduced by bioterrorists. I say "reintroduced" because we believe that smallpox has been eradicated in the

[6]Thanks to Dona Schneider for suggesting these.

world. The only known stores of smallpox virus are kept in two locations in the world, one in the U.S. and one in Russia. However, there is concern that terrorists could get hold of some of the virus or genetically engineer a copycat virus, and thus introduce a smallpox epidemic. One of the ways that the government seeks to get early warning about potential bioterrorist attacks with diseases like smallpox is to place networks of sensors/detectors to warn of such attacks. A key challenge is to determine where to best locate these sensors.

There are many issues surrounding the location of bioterrorism sensors that could make for good student activities. In 2006, DIMACS' work on epidemiology and bioterrorism, among other things, led to the center being named a "center of excellence" by the U.S. Department of Homeland Security (DHS). Through DyDAn, the DHS Center of Excellence for Dynamic Data Analysis that is based at DIMACS, I have already presented some potential activities to high school teachers that involve sensor location problems. (DyDAn is being supplanted by the new DHS Center of Excellence, the Command, Control, and Interoperability Center for Dynamic Data Analysis, CCICADA, also based at DIMACS.)

I first got involved with the bioterrorism sensor location initiative when the U.S. Defense Threat Reduction Agency asked me to think about the mathematical problem of locating sensors/detectors in an efficient way. I ended up learning a lot about the problem from the Institute for Defense Analyses and from the New York City Department of Health. Sensors are finding many uses in homeland security. We place sensors in buildings, on bridges, at border crossings, and even on uniforms of police. These sensors sense radiation, dangerous chemicals, biological agents, etc. Similar issues arise in placing sensors to protect against or give early warning about attacks with chemical or nuclear weapons or attacks on our networks: communication, financial, power, etc.

Bioterrorism sensor location problems are probably best explored in parallel between Biology and Mathematics classes. For example, Biology students could learn about how such sensors work. Typically, the sensors collect samples in the air over a period of time, like a day, and then those samples are brought to a laboratory for analysis. Students could study how the analysis is made and how long it takes, and think about how to speed up the process. Mathematics students could study more about the mathematics of efficient location patterns, which I describe below. In turn, Biology students could study the properties of the different types of "pathogens" of concern in bioterrorism and connect those properties up to the location problem.

Sensors are expensive. How do we select them and where do we place them to maximize "coverage," expedite an alarm, and keep the cost down? Approaches that improve upon existing, ad hoc location methods could save countless lives in the case of an attack and also money in capital and operational costs.

We can define two fundamental problems. The **Sensor Location Problem (SLP)** is to choose an appropriate mix of kinds of sensors and decide where to locate them for best protection and early warning. The **Pattern Interpretation Problem (PIP)** involves what to do when sensors set off an alarm. How can we help public health decision makers decide: Has an attack taken place? What additional monitoring is needed? What was the attack's extent and location? What is an appropriate response?

The first step in addressing the SLP is to formulate models for making it precise and measures of success of a sensor distribution plan. There are many possible formulations of the SLP and these present good explorations for your students in Math classes. What, indeed, is a measure of success? Is it to identify and ameliorate false alarms? To defend against a "worst case" attack or an "average case" attack? To minimize time to first alarm? (Worst case? Average case?) To maximize "coverage" of the area? To minimize geographical area not covered? To minimize size of population not covered? To minimize probability of missing an attack? Your students can be asked to come up with such criteria (and others). Other criteria involve cost. For example, given a mix of available sensors and a fixed budget, what mix will best accomplish our other goals? It is hard to separate the goals. Even a small number of sensors might detect an attack if there is no constraint on time to alarm. Without budgetary restrictions, a lot more can be accomplished. Of special interest are the biological characteristics of the different viruses that might be used in a bioterrorist attack. Among those of interest besides smallpox are plague and tularemia. They differ in size, weight, concentration needed to be dangerous, etc. In turn, these characteristics of viruses affect where to locate sensors that try to capture them. If viruses are disbursed through the air, lighter viruses might be higher up, for example, than lower ones. I can imagine some very interesting investigations for Biology students in connection with the location of sensors to capture viruses with particular characteristics.

One approach to the SLP is to develop new algorithmic methods. Developing new algorithms involves fundamental mathematical analysis. Analyzing how efficient algorithms are involves fundamental mathematical methods. Implementing the algorithms on a computer is often a separate problem – which needs to go hand in hand with the basic mathematics of algorithm development. All of these are topics appropriate for a Math class. For example, one can investigate "greedy algorithms." In such algorithms, we first find the "most important" location to place a sensor and locate a sensor there. (What things could "most important" mean?) Then we find the second-most important location. And so on. This approach has been explored by researchers at the Institute for Defense Analyses. It uses a "steepest ascent method" that doesn't guarantee an "optimal" or best solution but in practice gets close to optimal. Students could try this out by overlaying a map of New York City or Washington, DC with bioterrorism sensors, each of which has a circle around it representing area in which it can detect pathogens. One can see how few sensors can cover the region "almost" completely.

Problems of locating facilities (fire houses, garbage dumps, etc.) are classical problems in the field of the mathematical sciences known as operations research. Often, these problems are defined on networks with vertices and edges and with differing distances along edges. Users $u_1, u_2, \ldots, u_n$ are located at vertices. One approach is to locate the facility at vertex x chosen so that sum of distances to users is minimized. Thus, we want to minimize $\sum_i d(x, u_i)$, where $d(x, u_i)$ is the distance between x and u_i. Consider for example the network of Figure 14, where the vertices are places for users or facilities and every edge has the same distance, 1, as indicated. If $d(x, y)$ = length of shortest route from x to y, then, for example, $d(a, c) = 2$. Given users at $f = u_1, b = u_2$, and $c = u_3$, where do we place a facility to minimize the sum of distances to the users? A simple calculation shows that if $x = a$, then $\sum_i d(x, u_i) = 1 + 1 + 2 = 4$, while if $x = b$, then

$\sum_i d(x, u_i) = 2 + 0 + 1 = 3$. By calculating the same sums for other values of x, we find that $x = b$ is an optimal location for the facility. An alternative approach is to locate the facility at a vertex x chosen so that the maximum distance to one of the users is minimized. If $x = d$, then $\max_i d(x, u_i) = 2$ while if $x = e$, then $\max_i d(x, u_i) = 3$. By similar calculation, we find that with this notion of best location, there are three optimal locations, a, b, and d. Your students could solve similar location problems on other networks. (For a general reference on facility location problems, see for example Drezner and Hamacher [2004].)

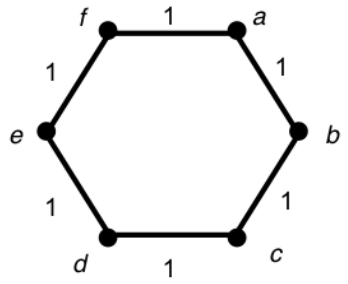

FIGURE 15. Given users at f, b, and c, the location that minimizes the sum of the distances from the location to the users is given by vertex b. The three locations that minimize the maximum distance to one of these users are a, b, and d.

This version of classical facility location is too simplified for bioterrorism sensor location. We don't have a network with vertices and edges; we have points in a city. Sensors can only be at certain locations (limited by size, weight, power source, hiding place, and perhaps by the biological and physical characteristics of the viruses or other pathogens being collected). We need to place more than one sensor. Instead of "users," we have places where potential attacks take place. Potential attacks take place with certain probabilities. Wind, buildings, mountains, etc. add complications. The type of biological agent used in an attack determines speed of spread, vertical and horizontal height above the ground, etc., etc. It would make for a good exercise to ask your students to suggest such complications.

The Pattern Interpretation Problem (PIP) can also lead to interesting classroom activities. It will be up to the decision maker to decide how to respond to an alarm from the sensor network. If a goal is to minimize false alarms, one approach is to use *redundancy*. We could require two or more sensors to make a detection before an alarm is considered confirmed. We could require the same sensor to register two alarms: The bioterrorism sensor known as "Portal Shield" requires two positives for the same agent during a specific time period. We could place two or more sensors at or near the same location and require two proximate sensors to give off an alarm before we consider it confirmed. Perhaps your students could suggest other ways in which redundancy could be used. Note that redundancy has drawbacks such as cost and delay in confirming an alarm. We need mathematical methods to analyze the tradeoff between lowered false alarm rate and extra cost/delay

Another approach to the PIP involves decision rules. Existing sensors come with a sensitivity level specified and sound an alarm when the number of particles collected is sufficiently high – above threshold. An alternative decision rule is to

sound an alarm if two sensors reach 90% of **threshold**, three reach 75% of threshold, etc. These rules could be formulated precisely in the language of mathematics and some examples illustrating such decision rules could make for interesting exercises. Biology students could collaborate with Mathematics students in designing these rules. For instance, could 75% of threshold have no biological significance while 90% does? Most likely, the answers to these questions are: We don't know or we think that 75% is dangerous with probability p while 90% is dangerous with probability $p' > p$. This will then lead to questions of the meaning of probability estimates and to risk assessment. Ah, but I could go on and on.

11. Closing Comments

Smallpox has taken me far afield. While my interest in and excitement about ecological models and DNA-RNA has continued, indeed grown, smallpox has gotten me into activities I never would have dreamed of less than a decade ago – using my background and skills to address problems of public health, the spread of disease, protection against bioterrorist attacks, climate change, and the cost of health care. There is a wide variety of topics that arise from the interconnections between the biological and mathematical sciences. So many of these are appropriate for high school students in both Biology and Mathematics classes. Introducing students to such topics should prepare them for a new world in which the traditional lines between disciplines are increasingly fuzzy, in which new careers are developing and new educational opportunities are opening up, in which it is incumbent upon us to understand Mathematics so as to understand Biology, and in which it is incumbent upon us to teach our students the value of Mathematics for so many applied problems that affect our world and their lives. Bringing the bio-math interface into the high schools is not going to be easy, but the rewards for doing so will be great.

References

[1] Adam, D., "Scientists create GM mosquitoes to fight malaria and save thousands of lives," *The Guardian*, http://www.guardian.co.uk/science/2005/oct/10/infectiousdiseases.medicineandhealth, Oct. 10, 2005.

[2] Anderson, R. M., and May, R. M., *Infectious Diseases of Humans*, Oxford University Press, UK, 1991.

[3] Benzer, S. "On the topology of the genetic fine structure," Proc. Nat. Acad. Sci. USA, 45 (1959), 1607-1620.

[4] Benzer, S., "The fine structure of the gene," Sci. Amer., 206 (1962), 70-84.

[5] Bernoulli, D., "De la mortalité causée par la petite vérole, et des avantages de l'inoculation pour la prévenir," Mém. Math. Phys. Acad. R. Sci. Paris (1760), 1– 45 (Imprimerie Royale; 1766); English translation: by L. Bradley, "An attempt at a new analysis of the mortality caused by smallpox and of the advantages of inoculation to prevent it," in L. Bradley (ed.), Smallpox Inoculation: An Eighteenth Century Mathematical Controversy, Adult Education Department, University of Nottingham, 1971.

[6] Blower, S.M., McLean, A.R., Porco, T.C., Small, P.M., Hopewell, P.C., Sanchez, M.A., and Moss, A.R. "The intrinsic transmission dynamics of tuberculosis epidemics," Nature Medicine, 1 (1995), 815-821.

[7] Board on Life Sciences, BIO2010: Transforming Undergraduate Education for Future Research Biologists, The National Academies Press, Washington, DC, 2003, http://www.nap.edu/books/0309085357/html/

[8] Dreyer, P.A., Jr., and Roberts, F.S., "Irreversible k-threshold processes: Graph-theoretical threshold models of the spread of disease and of opinion," Discrete Applied Mathematics, 157 (2009), 1615-1627.
[9] Drezner, Z, and Hamacher, H.W., Facility Location: Applications and Theory, Springer, 2004.
[10] Du, D-Z, and Hwang, F.K., Combinatorial Group Testing and its Applications, 2nd ed., World Scientific, Singapore, 2000.
[11] Fishburn, P.C., Interval Orders and Interval Graphs, Wiley, New York, 1985.
[12] Golumbic, M.C., Algorithmic Graph Theory and Perfect Graphs, Academic Press, New York, 1980.
[13] Hartnell, B., and Li, Q., "Firefighting on trees: How bad is the greedy algorithm?", Congressus Numerantium, 145 (2000), 187-192.
[14] Hastings, A., Arzberger, P., Bolker, B., Ives, T., Johnson, N., and Palmer, M., "Quantitative biology for the 21st Century," Report from an NSF-sponsored workshop on Quantitative Environmental and Integrative Biology held in December 2002 at the San Diego Supercomputer Center, http:/ /www.sdsc.edu/QEIB.
[15] Hastings, A., and Palmer, M.A., "A bright future for biologists and mathematicians," Science, 299 (2003), 2003-2004.
[16] Hethcote, H.W., and Yorke, J.A., Gonorrhea Transmission Dynamics and Control, Springer-Verlag, Berlin, 1984.
[17] Hinch, R., Greenstein, J.L., Tanskanen, A.J., Xu, L., and Winslow, R.L., "A simplified local control model of calcium-induced calcium release in cardiac ventricular myocytes," Biophys. J., 87 (2004), 3723-3736.
[18] Holley, R.W., Everett, G.A., Madison, J.T., Marquisee, M., and Zamir, A., "Structure of a ribonucleic acid," Science, 147 (1965), 1462-1465.
[19] Jackson, J.H., "Bioinformatics and genomics," in L.A. Steen (ed.), Math & Bio 2010: Linking Undergraduate Disciplines, Mathematical Association of America, 2005, 51-61.
[20] Kaplan, E.H., Craft, D.L., and Wein, L.M., "Analyzing bioterror response logistics: The case of smallpox," Mathematical Biosciences, 185 (2003), 33-72.
[21] Klein, E., Laxminarayan, R., Smith, D.L., and Gilligan, C.A., "Economic incentives and mathematical models of diseases," Environment and Development Economics, 12 (2007), 707-732.
[22] Lahav, G., Rosenfeld, N. Sigal, A., Geva-Zatorsky, N., Levine, A.J., Elowitz, M.B., and Alon, U., "Dynamics of the P53-Mdm2 feedback loop in individual cells," Nature Genetics, 36 (2004), 147-150
[23] Levin, S.A., Grenfell, B.., Hastings, A., and Perelson, A.S., "Mathematical and computational challenges in population biology and ecosystems science," Science, 275 (1997), 334-343.
[24] Lin, J., Andreasen, V., and Levin, S.A., "Dynamics of influenza A drift: the linear three-strain model," Math Biosci, 162 (1999), 33-51.
[25] MacGillivray, G., and Wang, P., "On the firefighter problem," J. Combin. Math. Combin. Comput., 47 (2003), 83-96.
[26] Morris, R.W., Bean, C.A., Farber, G.K., Gallahan, D., Hight-Walker, A.R., Liu, Y., Lyster, P.M., Peng, G.C.Y., Roberts, F.S., Twery, M., and Whitmarsh, J., "Digital biology: An emerging and promising discipline," Trends in Biotechnology, 23 (2005), 113-117.
[27] Mostashari, F., "Syndromic surveillance in New York City," Presentation to New York Academy of Medicine, Nov., 2002,
http://www.nyam.org/events/syndromicconference/presentationpdf/farzad_mostashari.pdf
[28] Palmer, M.A., Arzberger, P., Cohen, J.E., Hastings, A., Holt, R.D., Morse, J.L., Sumners, D., and Luthey-Schulten, Z., "Accelerating mathematical biological linkages: Report of a joint NSF -NIH workshop," February 2003, https://www.palmerlab.umd.edufreport.pdf
[29] Perelson, A.S., Neumann, A.U., Markowitz, M., Leonard, J.M., and Ho, D.D., "HIV-1 dynamics in vivo: Virion clearance rate infected cell life span and viral generation time," Science, 271 (1996), 1582-1586.
[30] Roberts, F.S., Discrete Mathematical Models, with Applications to Social, Biological, and Environmental Problems, Prentice-Hall, Englewood Cliffs, NJ, 1976.
[31] Roberts, F.S., Graph Theory and its Applications to Problems of Society, NSF-CBMS Monograph #29, SIAM Publications, Philadelphia, 1978.

[32] Roberts, F.S., Measurement Theory, with Applications to Decisionmaking, Utility, and the Social Sciences, Addison Wesley, Reading, MA, 1979. Digital printing, Cambridge University Press, 2009.
[33] Roberts, F.S., "Limitations on conclusions using scales of measurement," in S.M. Pollock, M.H. Rothkopf, and A. Barnett (eds.), Operations Research and the Public Sector, Handbooks in Operations Research and Management Science, Vol. 6, 1994, 621-671.
[34] Roberts, F.S., "Challenges for discrete mathematics and theoretical computer science in the defense against bioterrorism," in C. Castillo-Chavez and H.T. Banks (eds.), Mathematical and Modeling Approaches in Homeland Security, SIAM Frontiers in Applied Mathematics Series, 2003, 1-34.
[35] Roberts, F.S., "Meaningful and meaningless statements in epidemiology and public health," in B. Berglund, G. B. Rossi, J. Townsend and L. Pendrill (Eds.), Measurements with Persons, Taylor and Francis, 2010, to appear.
[36] Roberts, F.S., and Tesman, B., Applied Combinatorics, 2nd Edition, Chapman&Hall/ CRC, an imprint of Taylor&Francis, 2009.
[37] Sarkar, S., "Defining 'biodiversity': Assessing biodiversity," Monist., 85 (2002), 131-155.
[38] Setubal, J., and Meidanis, J. Introduction to Computational Molecular Biology, Brooks/Cole, Pacific Grove, CA, 1997.
[39] Steen, L.A. (ed.), Math & Bio 2010: Linking Undergraduate Disciplines, Mathematical Association of America, 2005, 51-61.
[40] Wang, P., and Moeller, S.A., "Fire control on graphs," J. Combin. Math. Combin. Comput., 41 (2002), 19-34.

DIMACS Rutgers University, 96 Frelinghuysen Road, Piscataway, NJ 08854-8018
E-mail address: froberts@dimacs.rutgers.edu