Are They Being Served?

A Beginning Mathematics Course for Students in the Biological Sciences

Carl Leinbach
Gettysburg College
Gettysburg, PA 17325
leinbach@gettysburg.edu

Abstract:

This paper intends to make a case for the use of a CAS (using a platform of either a PC with a CAS package such as Derive® or a hand-held CAS such as the TI-89® or Voyage 200®) in planning and teaching a mathematics course for students in the biological sciences. The course is to be constructed using the guidelines spelled out in the monograph BIO 2010 – Transforming Undergraduate Education for Future Research Biologists, [ ] published by the National Research Council of the National Academies in the United States. While this report does not formally suggest the use of a CAS, it does endorse the use of the technology that is available for the learning of tools and techniques that will aid students in the biological sciences. It also encourages an interdisciplinary approach to the construction and teaching of service courses.

A strong theme in this paper is that mathematicians need to look at the biological sciences and pay particular attention to the sweeping changes that are taking place in biological research and investigation. We need to tailor our courses to meet these needs. A simple 'watered down' mathematics course with exercises and examples that appear to use the biologists' language, but have little meaningful content does not really meet the biologists' needs. We will show how the use of a CAS can help to bridge the gap between the mathematics that is needed and the manipulations the students are prepared to do. The emphasis of the pedagogy is on content, meaning, and appropriateness of the technique, and not a drill on the mechanisms. In this paper examples are taken from the standard exponential growth as a starting point, population genetics as a means of analysis of the effect of natural selection, and, finally, a brief foray into the emerging important and mathematically rich field of bioinformatics.

Introduction:

Mathematics is an intrinsically beautiful and logically constructed abstract subject. Most of us love it because of these properties. We enjoy solving the many interesting puzzles that it presents to us. We are thrilled when we can make new discoveries within this system. However, for many of us, our initial attraction to mathematics was the ability to solve a particular problem or to gain new insight into an application. It is this ability to solve problems and provide insight that gives mathematics its premier position within the academic community. The reputation of mathematicians is that they are problem solvers. Unfortunately, it is this author's view that the mathematics we offer in our so called 'service courses' is not meeting the needs of all of our clients. This is particularly true in the biological sciences. Since the remarkable discoveries of Watson and Crick and the advent of high performance computing, biology has changed its research emphasis and also many of its research techniques. The discipline has embraced the information age and has added many significant areas of research. The standard "Calculus for Biologists" course provides neither the emphasis nor the depth that is needed for students to explore the frontiers of their discipline.

Calculus is an important subject for biologists, but their needs go far beyond the polynomial calculus clone of the mathematics majors' course that seems to be the norm. One of the primary interests of biologists is the need to understand change and how rates of change affect the entire

evolutionary and life processes. They need to be able to look at growth and extrapolate from their experimental data. In fact, one can argue that change is a central issue in modern biological research. Biology is becoming a quantitative/mathematical subject as opposed to its more classical descriptive role. The study of change and rates of change is central to understanding the mathematical nature of the subject and, thus explaining, the process.

This means that an important mathematical subject is an understanding of the derivative and the ability to solve differential equations. This goes far beyond the scope of the standard calculus course. Also, it calls for abilities that are beyond an untrained user of mathematics. This is where the CAS can play an important role. Derive and the TI hand-held CAS can solve most of the differential equations that the students in the biological sciences encounter. However, we now face a major dilemma. Have we reduced our beautiful subject to mere mindless button pushing and are we giving the students a "loaded gun" with which they can do more harm than good? The response to the first question is that we need to insure that the button pushing is informed and not mindless. The answer to the second question is yes we are giving them a loaded gun, and we have an obligation to teach a "fire arm's safety" course. In other words, we need to talk about the nature of the derivative, the information contained within the derivative, and how to check that our solution makes sense within the context of the problem. We need to examine differential equations and systems of differential equations for the information they contain about their solutions. We also need to discuss the sensitivity of the equations to the input parameters and the stability of the solution. If numerical techniques are used, we need to give the students insight into the appropriateness and stability of the technique. We need to train the students to ask the right questions and to develop a sense of skepticism about their results. Most importantly, we need to develop a confidence within the student that mathematics can address their important questions and build an attitude that will foster a confidence and cooperation between the two disciplines.

Our first example will consider the question of appropriateness of mathematical techniques for solving a basic and surprisingly rich mathematical idea, namely exponential growth. This is one of the basic equations of mathematical biology.

Growth of a Bacteria Population

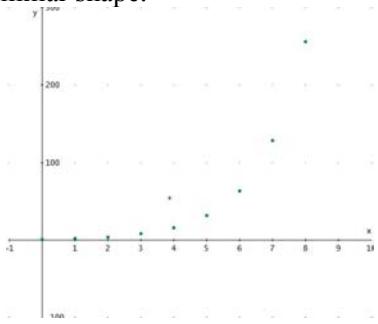
We start with a very basic problem.

> *A bacteria population has a present size of P_0 and it is observed to double in size every hour. What will be the size of the population after 5 hours and 40 minutes?*

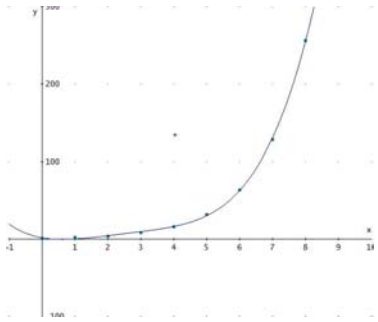
Let's consider the population for a period of 8 hours. We loose no generality in assuming $P_0=1$.

Hour	0	1	2	3	4	5	6	7	8
Population	1	2	4	8	16	32	65	128	256

If we plot this data we see the familiar shape.



Many students in a first college mathematics course are familiar with regression techniques. These techniques have found their way into most secondary school mathematics curricula. If we use Derive's **fit** function to fit these data with a fourth degree polynomial curve, we obtain the following graph.



This appears to be a very good fit! In fact, if we evaluate the polynomial at 5.67, we would obtain a value of about 49 which would compare favorably with experimental observation. However, *this solution makes no sense from the standpoint of the biological phenomenon!*

It is at this point that the biology and the mathematics come together to create the model for the basic growth. The biologist has observed that the growth of the population is proportional to the size of the population. In fact, the biologist has observed that the constant of proportionality is 2 per hour. Thus at time, *t*,

$$P_t = 2 * P_{t-1}$$

Now, the mathematics takes over.

$$P_t = 2 * P_{t-1} = 2 * 2 * P_{t-2} = 2 * 2 * 2 * P_{t-3} = \cdots = 2^t * P_0$$

The three dots in the above equations required a leap of faith, but they do open the door for a discussion of mathematical induction. More importantly, we have derived a general equation that is correct for whatever hour we choose. Also, we see that the 2 in the equation can be replaced by any constant of proportionality, λ. The mathematics has given more than was originally bargained for.

Now comes the really interesting question. What about time periods that are not exact hours? It is relatively easy to extend the idea to rational values of *t*. This would be enough to answer the question posed at the beginning of this section. However, we can use the problem of extending to all real values for *t* to begin a discussion of the calculus. The extension of the equation to all real values brings up the idea of continuity. Another approach can bring up the idea of linear approximation and the derivative. One can even mention and begin an examination of the basic differential equation,

$$\frac{dP}{dt} = \lambda * P$$

Of course, at this stage the solution will be premature without a discussion of the exponential function.

The moral of the story is clear. The biology has lead us to an interesting mathematical excursion that will involve interested biology students and score points for the importance of mathematics

and mathematical analysis within their discipline. This is far superior to allocating the biology to one example in the course of a mathematical development.

The next example will show the use of difference and differential equations in the course of the analysis of an evolutionary process.

Hardy-Weinberg Equilibrium and Natural Selection

We will consider a one locus, two allele genetic trait among a large population that features distinct generations. We will also assume random mating within the population. These assumptions can apply to many fish, coral, and plant populations. We will assume that the timing between generations is as follows:

Generation	Zygote		Adult	Gametes		Zygote	Generation
t	•	$\rightarrow$	*Breeding*	o ~	$\rightarrow$	•	$t+1$

The arrows indicate two differential survival points. The first is from the zygotic stage to adulthood and the breeding stage. The second is number of gametes that form the zygotes for the next generation. We combine these two into a single survival or fitness rate. If we are considering a two allele genetic trait with dominant, *d*, and recessive, *r*, characteristics we denote the survival (or fitness) rates for the traits among the zygotes as W_{dd}, W_{dr}, and W_{rr}. Note that the pairings *dr* and *rd* result in the same class of zygote. If allele *d* is distributed among the population with frequency, *p*, and allele *r* is distributed among the population with frequency, *q*, and then the following facts hold.

$$p + q = 1$$
$$p^2 + 2pq + q^2 = (p+q)^2 = 1$$

The second relationship is important algebraically and also in light of the fact of random mating. p^2 is the probability of a *dd* mating, *pq* a *dr* mating, and q^2 an *rr* mating. The total number of alleles in the next generation of the population is given by the difference equation:

$$N_{t+1} = (p_t^2 W_{dd} + 2p_t q_t W_{dr} + q_t^2 W_{rr}) N_t$$

The number of *d* alleles is:

$$N_{d,t+1} = (p_t^2 W_{dd} + p_t q_t W_{dr}) N_{d,t}$$

The number of *r* alleles is:

$$N_{r,t+1} = (q_t^2 W_{rr} + p_t q_t W_{dr}) N_{r,t}$$

The difference equations for the associated frequencies are:

$$p_{t+1} = \frac{(p_t W_{dd} + q_t W_{dr}) p_t}{p_t^2 W_{dd} + 2p_t q_t W_{dr} + q_t^2 Wrr}$$
$$q_{t+1} = 1 - p_{t+1}$$

Doing some algebraic manipulation keeping in mind that for any *t*, $q_t = 1 - p_t$, we approximate this system of difference equations with the system of differential equations:

$$\frac{dp}{dt} = \frac{pq[p(W_{dd} - W_{dr}) - q(W_{rr} - W_{dr})]}{p^2 W_{dd} + 2pqW_{dr} + q^2 W_{rr}}$$
$$\frac{dq}{dt} = -\frac{dp}{dt}$$

We have used our mathematical analysis to predict the allele frequencies within a general population. However, we need to see how these equations relate to the biological intuition and data that our students possess. Up to this point our mathematical manipulations have been just

that, mathematical manipulations. Let's use our CAS to display the effects of environmental conditions on the allele frequencies. In essence, we will be looking at the effects of natural selection.

In the following a simple Derive program was written to evaluate the system of difference equations for n generations. We will show a graphical representation of the results of the evaluations. However, prior to this we consider the case where there is no selection process going on.

No Natural Selection

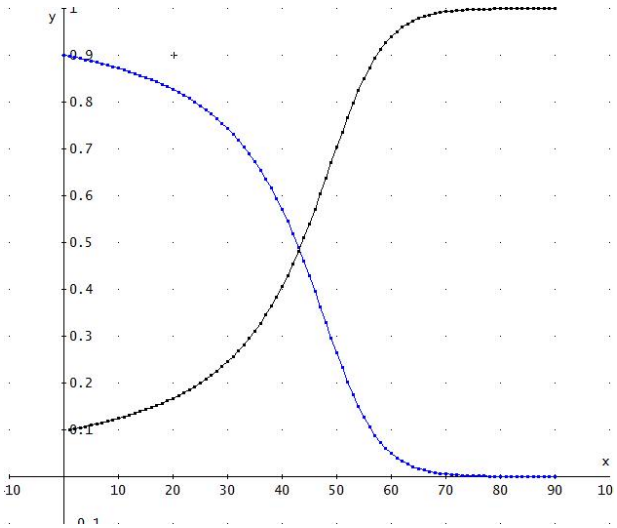
In this case all zygotes have equal survival rates, i.e. $W_{dd} = W_{dr} = W_{rr}$. Then since $p + q = 1$, we have

$$p_{t+1} = p_t \quad \text{and} \quad q_{t+1} = q_t$$

for all t. This means that the system is in an absolute equilibrium and the frequencies remain constant throughout time.

Selection against the dominant allele

Since it is the relative frequency of the W's that matters in our difference and differential equations and not the population size, we will assume that the W's are values between 0 and 1. In this case we will choose $W_{dd} = W_{dr} = .8$ and $W_{rr} = 1$. This difference between the survival values, although small, will prove to be deadly to the dominant allele. In the following graph we see that within about 70 generations the dominant allele is all but extinct within the population.

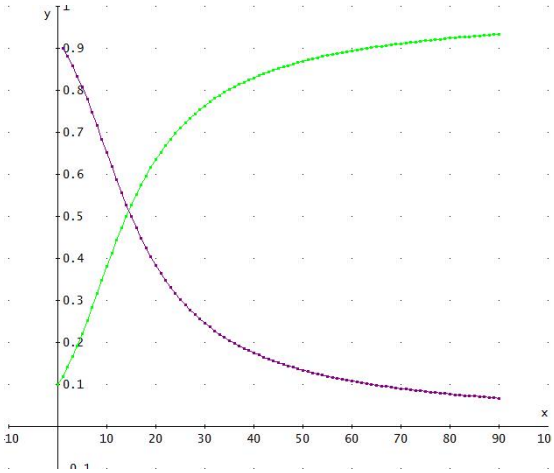


Even if the original difference in the survival rate is very small, the effect is the same: the dominant allele dies out. It may take thousands of generations, but eventually an equilibrium state is achieved at the expense of the dominant allele. Notice the shape of the curve. Because of the initial superiority of p over q, the dominant allele while decreasing in frequency appears to be resisting the effects of selection. However, as q increases the rate of decrease for p accelerates. This effect can be predicted by the differential equation form of the selection equations.

Finally, note that the shape of the curve for q is essentially a reflection of the shape of the curve for p.

Selection against the recessive allele

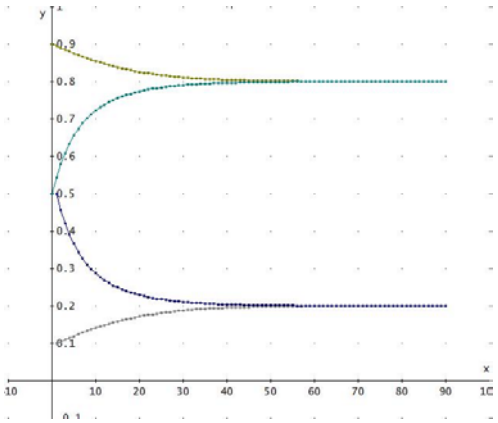
In this case the shoe is on the other foot. We will choose $W_{rr} = W_{dr} = .8$ and $W_{dd} = 1$. Some interesting questions for our biology students are: What assumptions are we making when we set the survival rate of the heterozygote to be the same as that of the homozygote that is being selected against? Do we need to make these assumptions? Do they make sense from a biological standpoint? We will give the recessive allele an edge by assuming that it is the most common allele in the population.



While the recessive allele is not quite extinct after 90 generations, it is headed for extinction. What is interesting is that the shape of the curve is different than that in the previous case. Here the initial descent is very sharp and the curve flattens out as the allele nears extinction. Can this shape be predicted from the differential equation form of the system of equations for the allele frequencies?

Selection in favor of the Heterozygote

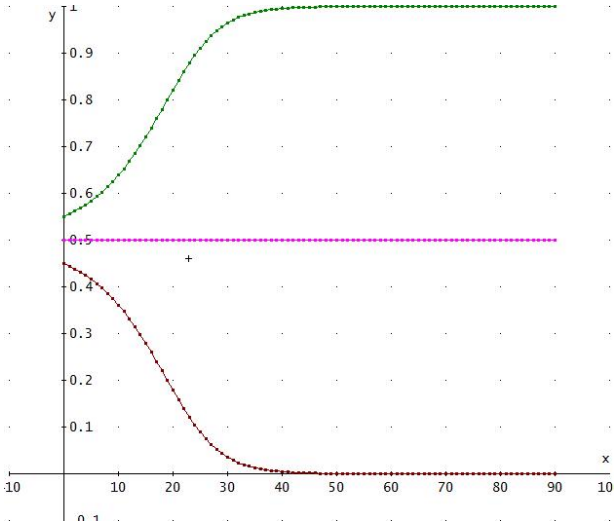
This case is more interesting than our previous two cases and the results are less to be expected. We will assume that $W_{de} = 1$. We will not set the selection against the remaining two homozygotes to be equal. We will set the selection against the recessive homozygote $(1 - W_{rr})$ to be four times as great as the selection against the dominant homozygote. Thus, we set $W_{dd} = .9$ and $W_{rr} = .6$



In this graph we considered two examples. In the first, p was set to its usual value of .9 and q= .1. In the second we set p and q both equal to .5. In both cases we see that p eventually tends to .8 and q tends to .2. This opens the door to a discussion of equilibrium situations. These values can be attained by setting the derivatives in the differential equation form of our equations equal to 0.

Selection against the Heterozygote

This case occurs very rarely in nature. However, it is interesting to consider. We will consider $W_{dd} = W_{rr} = 1$ and $W_{dr} = .8$. We will only graph the value of p since the result for q will be clear. The upper graph is typical of the graph for any choice of a value for p above .5. The lower graph is typical for any choice of p less than .5. The middle graph is, of course the result of choosing p equal to .5.



In this case we have an unstable equilibrium. The value of .5 is an equilibrium point, but it is repelling for all values of p other than .5. We see that any movement away from .5 will result in the elimination of one allele from the population.

Conclusion

In each of the individual cases above, it is possible to substitute our numerical values for the W's and use our CAS to obtain analytical solutions. In fact, it may be a useful exercise to use the CAS in this way. On the other hand, our students will gain little insight into the overall evolutionary process over what they can see using the graphs and analyzing the derivative of the frequencies in each case. The answer is not always the answer. It is the process and an understanding of the derivative that provides the insight.

The numerical technique that was used to solve the equation is among the simplest possible. However, it is representative of the assumptions that we made for our population, namely, that there are distinct generations. In this case, the differential equation is the approximation to the actual data. In other cases, the differential equation may be more representative of the biological process.

The Biology of the Future – Bioinformatics

The previous sections were closely linked to the relationship of biology to analysis of difference and differential equations. In this section we will look at the emerging field of bioinformatics. In one sense, it is more descriptive than analytical. However, there are rich applications in terms of algorithmics, pattern matching, and statistical analysis. These subjects all have their roots in mathematics. A course for biology majors should include an introduction to the techniques of DNA analysis and string comparisons.

The advent of computing power coupled with the advances in molecular biology has opened new and exciting areas for biologists. It also provides opportunities for mathematics departments while teaching students in the biological sciences. In order to understand modern biology students need to be able to generate hypotheses, formulate algorithms for searching large databases, evaluate the efficiency of the algorithms, and compare their results with real data. Many topics and techniques from discrete mathematics apply to the conduct of good research in modern biology. The appropriate and efficient use of technology provides an opportunity to do solid analysis.

This paper will not delve into all of the possibilities that are available in this rich field of investigation. It will illustrate and apply a programmable CAS for the alignment of two, relatively short DNA sequences. This investigation will illustrate the algorithms that are working in the major tools that are available to research biologists through services offered on the web via PubMed and other similar sites. In the United States these resources are available through the National Center for Biological Information web site which is linked to several web sites throughout the world.

DNA with its well known double helical structure consists of four nucleotides: Adenine (A), Cytosine (C), Guanine (G), and Thymine (T). These nucleotides bind together according to certain rules to form chains that contain genes and the basic building blocks of life, itself. A fundamental problem for researchers is to compare a sequence they may have encountered with the database of known DNA sequences.

Sequence comparison is far from simple. First of all, where in the given sequence does one begin the comparison? There is no clear cut beginning or end to such a sequence. The sequence can change as a result of mutation, genetic drift, insertions, or deletions. How do we account for these changes and still recognize similarities? The combinatorial possibilities are overwhelming!

In the following we will use Derive to illustrate some of the basic techniques of simple sequence comparison. The first will be a simple matching of two sequences.

Dot Plots

We begin with two simple nucleotide sequences. They are defined as Derive variables, s1 and s2.

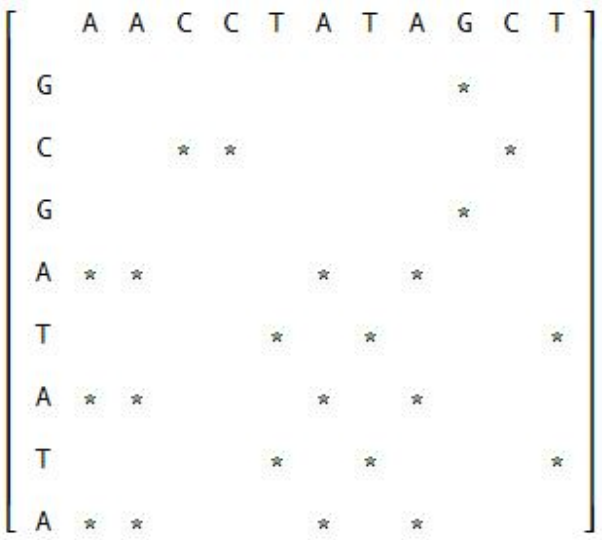
s1 := AACCTATAGCT

s2 := GCGATATA

The question that we will address is whether these sequences share any common subsequences and can they be aligned in a way to show any similarities?

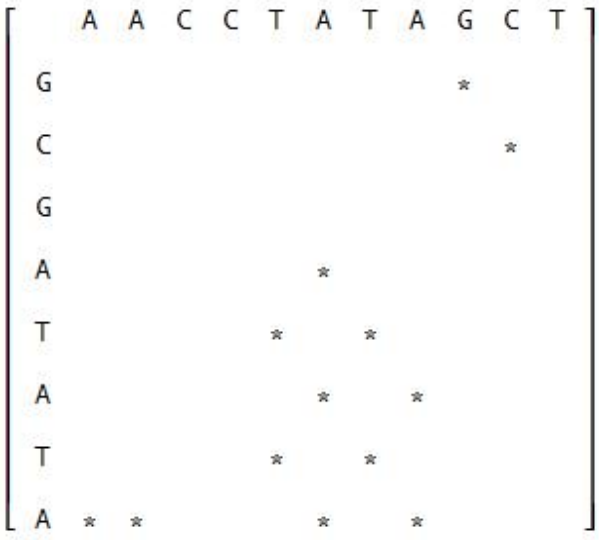
Our first strategy is to create what is known as a Dot Plot. This is merely a data representation of the two sequences as a matrix. The sequence, s1, is used to label the columns and s2 labels the rows. A dot or star is placed in cell i,j if the i-th letter of s2 matches the j-th letter of s1. Otherwise, the cell is left blank. One mathematical benefit of this technique is that students learn about matrices and setting the contents of the various cells within a matrix.

This strategy can be implemented in a Derive program to produce the following matrix. This program is straight forward and is easily explained to the students. Viewing the program further reinforces the students' understanding of the layout of a matrix.



If one carefully observes this matrix, the user can observe some patterns of *'s along certain forward diagonal sequences. These sequences mean that the corresponding subsequences of the two strings match up. However, we are dealing with two unrealistically short strings of DNA in this example. Longer sequences would produce much more cluttered Dot Plots. We need some strategy to clean up the display so that the matching subsequences will be more apparent.

One simple strategy is to display only those positions where at least one of the adjacent nucleotides matches the nucleotide in the other sequence. This requires a modification of our Derive program. It also requires an understanding of some basic logical inference in order to construct the decision statements to implement the strategy. The Derive program for this is included in the author's DFW file entitled *Bioinformatics.dfw*. The new matrix is shown below



In this matrix the matching subsequence TATA stands out more prominently. It is the longest matching subsequence, however there is another possibility, ATA. We choose a match of the form:

A A C C T A T A G C T
G C G A T A T A - - -

But, exactly how good is this match? There are other possibilities.

The –'s in the above representation of s2 are called "gaps." They indicate an insertion in the other sequence, s1, or a deletion from s2. The common term for this is an "indel". The inclusion of gaps in a sequence further complicates our matching problem. The simple inclusion of three gaps in s2 increases the number of possible target sequences to $C(9, 3) = 220$ possible matches. If we allow gaps in both sequences can expand the possibilities to the hundreds of thousands. For example we could propose a match of the form.

- A A C C - T A T A G C T
G - - C G A T A T A - - -

Is this a better match than the above? Also, how does one discover such matches? Obviously Dot Plots and Refined Dot Plots can take us only so far in our quest for a sequence match.

One fact to keep in mind is that indels, although found quite often in DNA sequences are not a common biological occurrence. They are a 'mistake'. It is just that we are dealing with billions of DNA strands for even the simplest of biological specimens. So, even low frequency events can be found. We compensate for this by allowing gaps, but establishing a penalty for them in the overall scoring of sequence match.

Scoring Matrices and More Sophisticated Matching Techniques

By how much should an exact matching of two nucleotides benefit an alignment? What penalty should be assessed for a mismatch? A gap? These are questions that have quantitative answers, but no exact answers. It is necessary to rely on our colleagues in the biological sciences for answers. The answers are based on heuristics and there are more than one set. When these questions are decided upon, then we construct a scoring matrix. The popular BLAST (Basic Local Alignment Search Tool) scoring matrix assigns a score of 5 to a match and penalties of -4 for a mismatch and -8 for a gap. In this presentation, another popular scoring scheme is used. A match is assigned a score of 1, a mismatch is given a score of 0, and a gap is penalized with a -1. The scoring of DNA sequences is relatively simple. The situation for Protein sequences is more complex due to the fact that there are 20 Amino Acids and nature freely substitutes one for another in some cases. This discussion will deal with DNA and the following scoring scheme

	A	C	G	T
A	1	0	0	0
C	0	1	0	0
G	0	0	1	0
T	0	0	0	1

with a gap penalty of -1.

The algorithm for constructing the score for the alignment of two DNA sequences using this scheme is due to Needleman and Wunsch. It is a dynamic programming algorithm. It is described at the top of the next page.

Needleman-Wunsch Algorithm

0. Lay out the matrix with s1 across the top and s2 down the left. The dimension of the matrix is (DIM(s2)+1) *X* (Dim(s1) +1).
1. Place –(i – 1) in cell i of row 1 and –(j – 1) in cell j of column 1.
2. Starting in cell (i, i) with i ≥ 2, compute the following three values:
    a. The value in the adjacent cell to the left minus 1
    b. The value in the adjacent cell above minus 1
    c. The value in the cell diagonally above the cell to the left plus 0 if the cell represents a mismatch and plus 1 if the cell represents a match.
3. Choose the maximum of these three values and place it in the cell. Note which cell was chosen for the computation of the value in the cell.
4. Repeat 2 a, b, and c and 3 above for all of the cells remaining in row i and column i.
5. Repeat steps 2, 3 and 4 until all of the cells of the matrix are evaluated.
6. The value of the cell in the lower right corner is the score of the alignment. The alignment can be retraced starting in this cell and moving in the direction indicated by the present cell. A diagonal move indicates an alignment of the two nucleotides represented by the cell. A vertical move indicates a gap in s1 and a left move a gap in s2.

Instead of applying this algorithm to s1 and s2 given above, we apply it to two slightly more interesting sequences.

s3 ≔ ACTCG

s4 ≔ ACAGTAG

$$\left[\begin{bmatrix} 0 & -1 & -2 & -3 & -4 & -5 \\ -1 & 1 & 0 & -1 & -2 & -3 \\ -2 & 0 & 2 & 1 & 0 & -1 \\ -3 & -1 & 1 & 2 & 1 & 0 \\ -4 & -2 & 0 & 1 & 2 & 2 \\ -5 & -3 & -1 & 1 & 1 & 2 \\ -6 & -4 & -2 & 0 & 1 & 1 \\ -7 & -5 & -3 & -1 & 0 & 2 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 1 & 1 & 1 & 1 \\ 0 & 3 & 2 & 1 & 2 & 1 \\ 0 & 2 & 3 & 2 & 2 & 2 \\ 0 & 3 & 3 & 2 & 2 & 2 \\ 0 & 3 & 3 & 2 & 2 & 2 \\ 0 & 2 & 3 & 3 & 2 & 2 \\ 0 & 3 & 3 & 3 & 2 & 2 \end{bmatrix}, \begin{bmatrix} \text{AC--TCG} \\ \text{ACAGTAG} \end{bmatrix}\right]$$

In the interest of saving space the sequences were not written across the top and along the left of the matrices. The left hand matrix is the matrix of scores. The score for the alignment on the right is 2. Is this a good score? That needs to be tested. One popular method is to align the first against several rearrangements of the second using this algorithm. Then note the median and standard deviation of the resulting scores. If the score of the present realignment is far enough (2 or more standard deviations) above the median, we suspect that we have a good alignment.

The second matrix tells the direction that determined the score in each of the cells. A 1 indicates that the score was determined by the cell to the left. A 2 indicates the cell diagonally above, and a 3 the cell directly above.

The third matrix is the alignment that results from starting in the lower right cell and generating the alignment in reverse by following the directions indicated by the second matrix.

This figure was generated using the Derive program NWAlign that is found in the DFW file, *Bioinfomatics.dfw* that was submitted with this paper.

The Semi-Global Alignment Algorithm

The Needleman-Wunsch Algorithm appeared to do a fine job on the sequences given above; however, consider the following example from our Derive file.

(NWAlign(ACGT, AAACACGTGTCT)) $_3$

$$\begin{bmatrix} \text{----ACG-T} \\ \text{AAACACGTGTCT} \end{bmatrix}$$

This is clearly not a good alignment! The first sequence is an exact subsequence of the second. The problem is that the Needleman-Wunsch Algorithm penalizes gaps at either end of the sequence at the same rate that it penalizes gaps in the middle. The gaps in the middle are caused by indels. The gaps at either end may be due to incomplete sampling.

The algorithm needs to be modified to correct this problem. The obvious cause of the problem is in step 1 of the Needleman-Wunsch Algorithm. Instead of applying the gap penalty of -1 to leading and trailing gaps, treat them as mismatches and assign a cell value of 0 to all of the cells in row 1 and column 1. This strategy is implemented in the Derive program SGAlign in the *Bioinformatics.dfw* file. The following is the result.

$$\left[\begin{bmatrix} 0&0&0&0&0 \\ 0&1&0&0&0 \\ 0&1&1&0&0 \\ 0&1&1&1&0 \\ 0&0&2&1&1 \\ 0&1&1&2&1 \\ 0&0&2&1&2 \\ 0&0&1&3&2 \\ 0&0&0&2&4 \\ 0&0&0&1&4 \\ 0&0&0&0&4 \\ 0&0&1&0&4 \\ 0&0&0&1&4 \end{bmatrix}, \begin{bmatrix} 0&0&0&0&0 \\ 0&2&2&2&2 \\ 0&2&2&2&2 \\ 0&2&2&2&2 \\ 0&2&2&2&2 \\ 0&2&3&2&2 \\ 0&2&2&2&2 \\ 0&2&3&2&3 \\ 0&2&2&3&2 \\ 0&2&2&2&3 \\ 0&2&2&2&3 \\ 0&2&2&2&3 \\ 0&2&2&2&3 \end{bmatrix}, \begin{bmatrix} \text{----ACGT----} \\ \text{AAACACGTGTCT} \end{bmatrix}\right]$$

This is obviously the correct alignment. The score for this alignment is 4. If one checks out the score for the Needleman-Wunsch algorithm the score is –4.

Conclusion

We have illustrated only the barest beginnings of bioinformatics. It should be apparent that there are rich areas for mathematical discussion and the application of mathematics and mathematical reasoning. As the area grows more and more opportunities will arise for mathematics. This means that we will need to be constantly evaluating what we teach in our mathematics for students in the biological sciences. In the above we saw that an understanding of matrices was essential. We used the dynamic programming algorithm to efficiently reduce the order of the computations involved in aligning two sequences. We did not discuss the analysis of these algorithms and the general topic of Analysis of Algorithms. Certainly, this subject deserves coverage in a mathematics course for biologists.

The structure of protein sequences is a very complex subject. Sequence alignment deals with a linear structure. Protein sequences have a 3-D representation. Looking at their sequencing is not sufficient to completely describe the Protein. It is necessary to determine where the protein "folds." This is a difficult question and it requires the cooperation of mathematicians, computer scientists, biologists, and bio-chemists. We have a rich source of research problems and pedagogical opportunities in our future.

Summary

In Bio2010 the Computer Science and Mathematics Panel makes the following statement: "Rather than doing the standard calculus, linear algebra, and differential equations, a one year course on mathematics for biologists should be designed. This course should be based on biological examples and include methods of solving problems, but with more emphasis on standard packages, …, than a course for mathematics majors …" [5, p169] This paper was an attempt to echo that thought and give some examples for course content. As was stated earlier: the answer is not always THE answer. Many times the answer lies in constructing the model and knowing what the model can tell about the biological process.

Bibliography

1. Attwood, T.K and Parry-Smith, D.J., Introduction to Bioinformatics, 1999, Pearson Education Limited, Harlow, England.
2. Claverie, J-M. and Notredame, C., Bioinformatics for Dummies, 2003, J. W. Wiley Inc., New York, NY.
3. Hoppensteadt, F. C. and Peskin, C. S., Mathematics in Medicine and the Life Sciences, 1992, Springer-Verlag New York Inc., New York, NY.
4. Krane, D. E. and Raymer M. L., Fundamental Concepts of Bioinformatics, 2003, Benjamin-Cummings, San Francisco, CA.
5. National Research Council, BIO2010, Transforming Undergraduate Education for Future Research Biologists, National Academies Press, Washington, DC
6. Roughgarden, J. , Theory of Population Genetics and Evolutionary Ecology: An Introduction, 1979 (republished 1996), Prentice-Hall, Upper Saddle River, NJ