CONTS

# Are They Being Served?

## A Proposal for a Beginning Mathematics Course for Students in the Biological Sciences

Carl Leinbach
Gettysburg College
Gettysburg, Pennsylvania, USA
leinbach@gettysburg.edu

# Who are our Students?

- Students in the Biological Sciences come to the Mathematics Department for tools to give insight about Biological Processes
- They also want tools to help them solve their problems and to understand questions that arise in their discipline
- Some may even see the need to take a Mathematics Course as a hurdle that they must clear in order to gain entry into their chosen field.

# The Mathematics Department's Standard Response

- No Special Course for Biology Majors. Students placed either in Mathematics Majors course or a general 'Applied Calculus' course.
- Applications are shown in problems or short examples and are generally elementary problems from the area or artificial problems using jargon. Little time spent motivating the problem within the discipline's context.
- We solve the problems we want to solve, not the ones that they want solved.

# What is the Net Result?

CONTS

- Students generally fail to see the value of taking a mathematics course as a requirement for entry into their major
- The many new areas for applications of Mathematics in Biology and Biological Research are virtually ignored by the Mathematics Department. It also gives very little attention to existing applications in areas such as genetics, population biology, and ecology.
- We may, in fact, be doing a disservice to an area that requires a great deal of assistance from mathematics

CONTS

# A Call for Action

BIO2010 – Transforming Undergraduate Education for Future Research Biologists, National Research Council of the National Academies, The National Academies Press, Washington, D.C. 2003 www.nap.edu

". . . it is important that all students understand the growing relevance of quantitative science in addressing life science questions."

"… faculty should attempt to utilize teaching approaches that are most likely to help students learn these skills."

# Skills That Are Important

- Understanding rates of change and growth
    1. Difference Equations
    2. Differential Calculus
        - Emphasis on Meaning
        - Minimal Time on Calculation
        - Ability to build and interpret models
- Ability to use tools to solve Differential Equations and interpret models and solutions
    1. Understanding of issues such as equilibrium, sensitivity, and stability.
    2. Understand the appropriate use of symbolic and numerical tools.
    3. Ability to check and interpret symbolic and graphical solutions

# Skills That Are Important (continued)

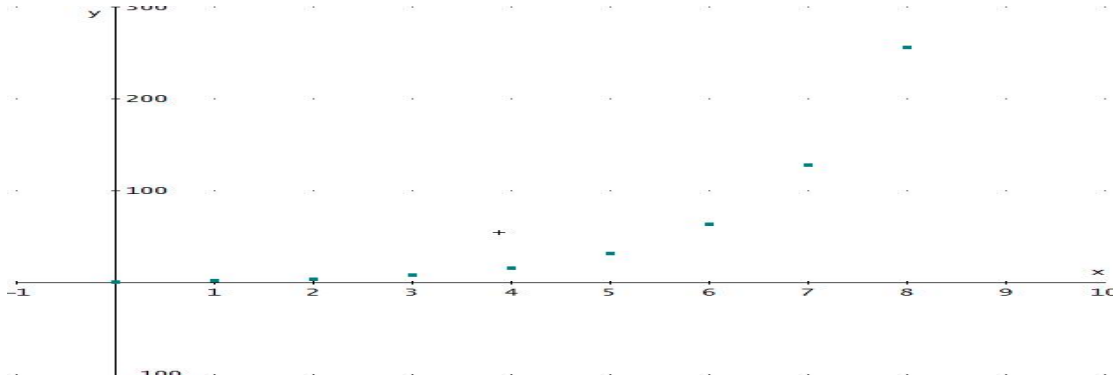
- Interpret the meaning of the Definite Integral within their context
    1. Population Size and Totals.
    2. Statistical Interpretations
    3. Appropriate Use of CAS Tools
- Develop mathematical reasoning abilities to interpret and solve the problems of modern biological research
    1. Logical Inference
    2. Development of Algorithms
    3. Analysis of Algorithms
    4. Ability to access large databases

# A Simple, Basic Example

## Exponential Growth

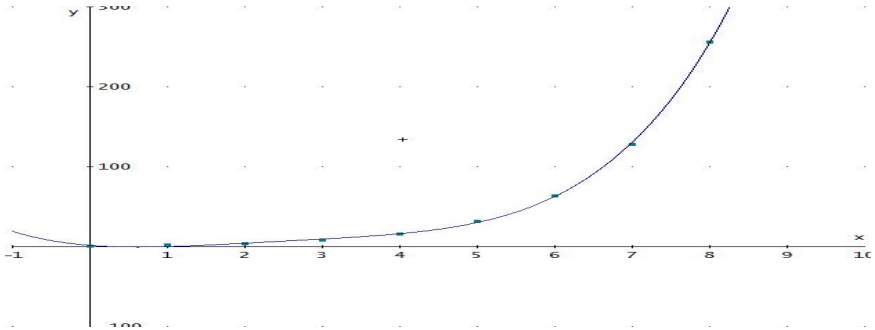
- Well Known, Covered in Virtually Every Calculus Course

| Hour | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Popultion | 1 | 2 | 4 | 8 | 16 | 32 | 64 | 128 |



# Exponential Growth (continued)

- Fitting the data with a fourth degree polynomial
- Rather good fit,  gives answers within acceptable error range
- Totally Inappropriate!  Ignores the basic, underlying process



# Exponential Growth (continued)

The Process:

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

Leads to a discussion of the general exponential function, logarithms, and the differential equation:

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

A

Good Biology $\rightarrow$ Good Mathematics

c

o

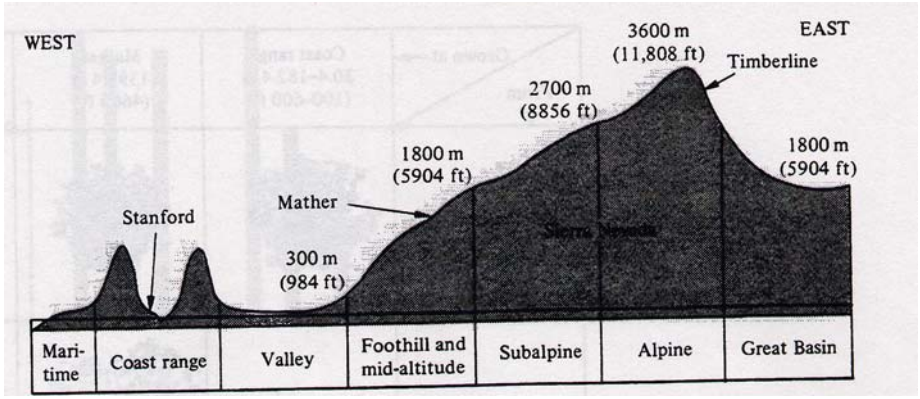
m

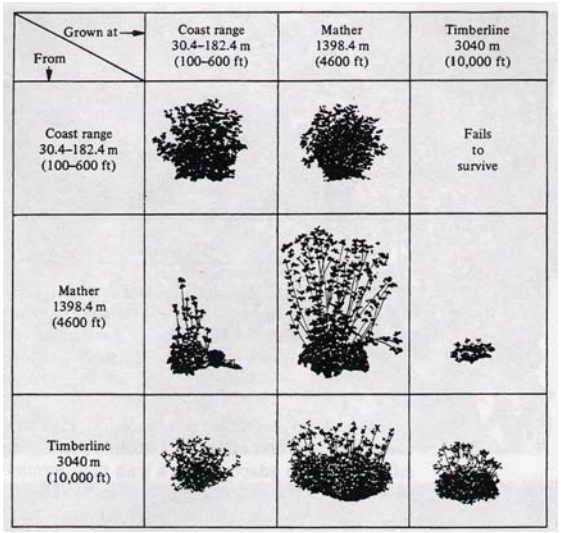
CONTS

# Hardy – Weinberg Equilibrium & Natural Selection

| Generation | Zygote | | Adult | Gametes | | Zygote | Generation |
|---|---|---|---|---|---|---|---|
| $t$ | • | → | *Breeding* | o ~ | → | • | $t + 1$ |

Survival Rates Influenced By Selection: $W_{dd}$ , $W_{dr}$ , $W_{rr}$



We consider a one locus, 2 allele, large population that features distinct generations

$p$ = frequency of dominant allele

$q$ = frequency of recessive allele

Then, $p + q = 1$

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

Total number of alleles

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

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

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

# Important System of Difference Equations

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

$$q_{t+1} = 1 - p_{t+1}$$

Leading to the System of Differential Equations

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

$$\frac{dq}{dt} = -\frac{dp}{dt}$$

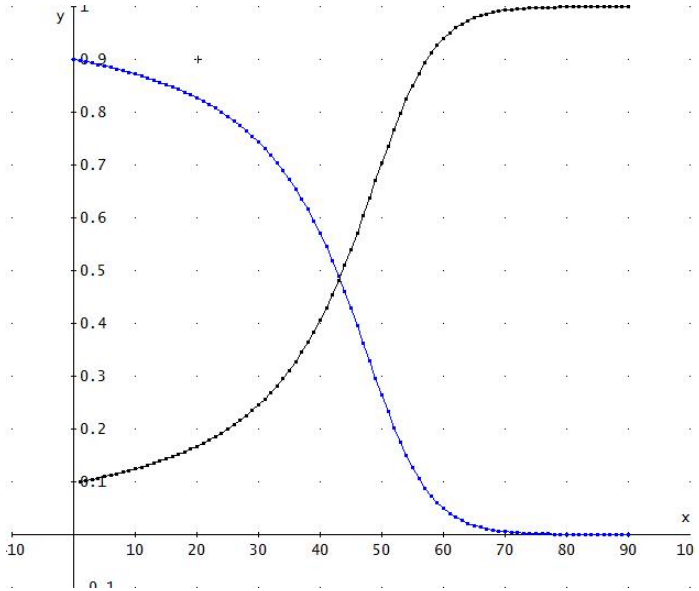
# What do they tell us?

Basic Case: Wdd = Wdr = Wrr

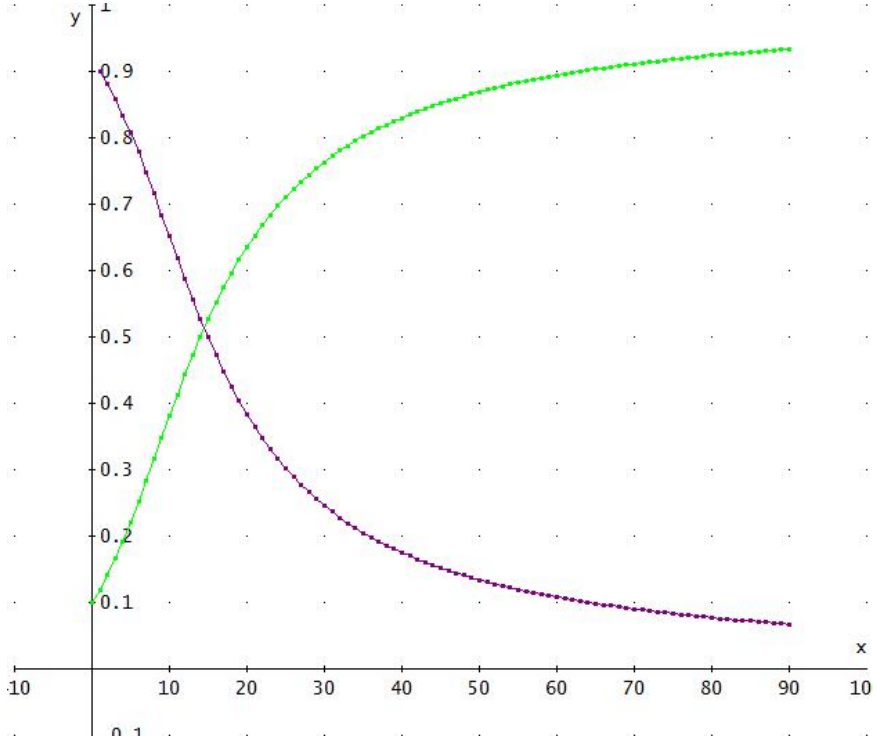
$p_{t+1} = p_t$ and $q_{t+1} = q_t$ for all $t$

Now the interesting cases

1. Selection against a dominant allele $W_{dd} = W_{rr} = 1$ and $W_{dr} = .8$.



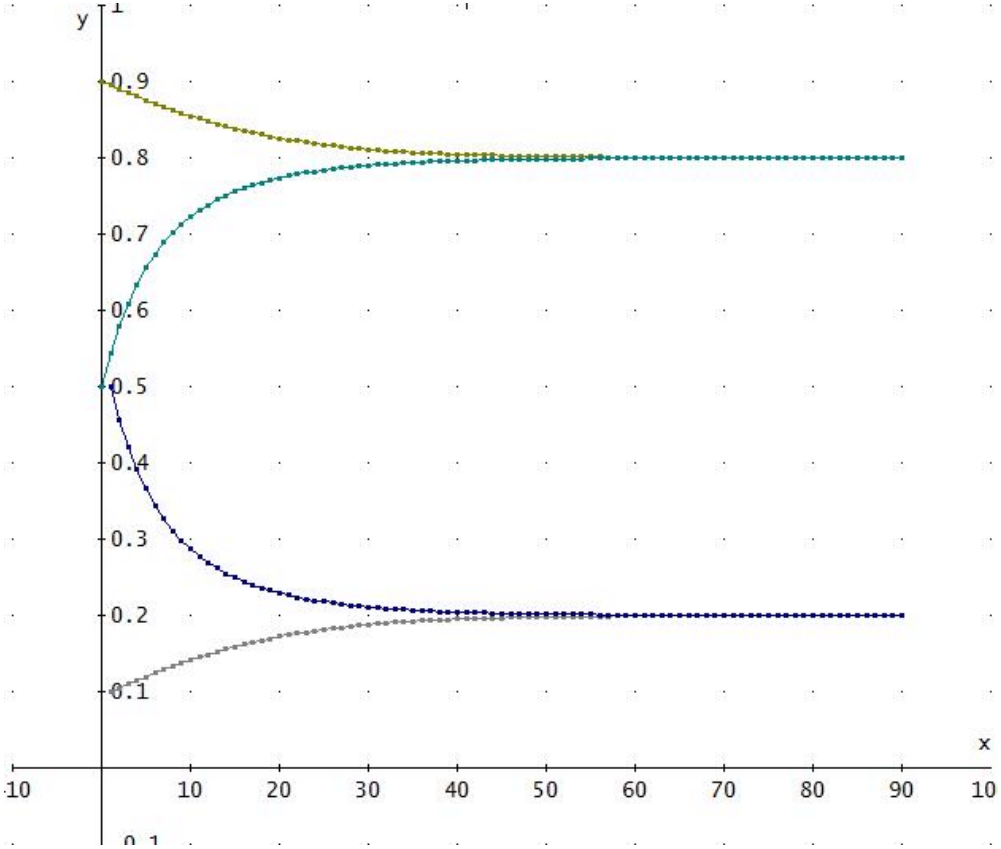
2. Selection against a recessive allele $W_{rr} = W_{dr} = .8$ and $W_{dd} = 1$.




Note the difference in the shape of the curves between this and the previous case.

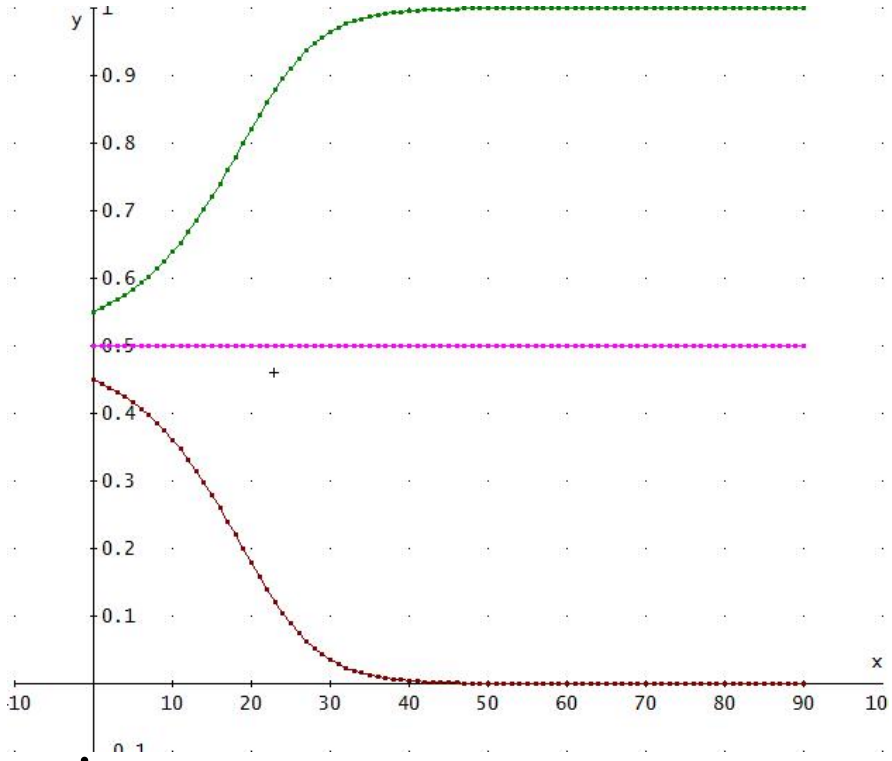
3. Selection in favor of the heterozygote $W_{rr} = .6$

$W_{dr} = 1$ and $W_{dd} = .9$.




Note the existence of stable equilibria

4. Selection against the heterozygote $W_{dd} = W_{rr} = 1$ and $W_{dr} = .8$




Very rare in nature.

Note: The existence of an unstable equilibrium that leads to the elimination of one allele

# Conclusions about Calculus

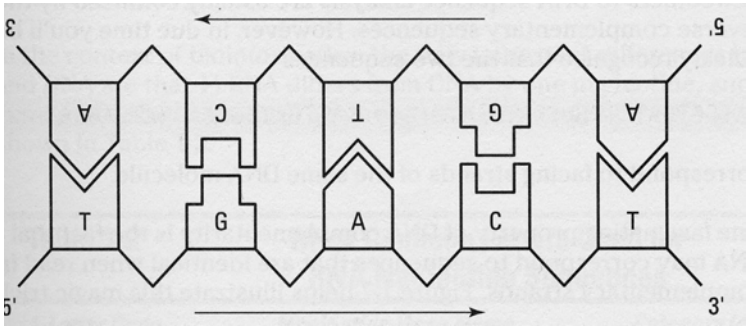
- The answer is not always THE answer.
- Knowing about the derivative is as important as knowing how to calculate the derivative, in fact, for these students it is more important.
- Need to know enough about the calculations to be able to trust and understand the results of the CAS.
- Using a CAS to solve differential equations either analytically or numerically may be giving students a “loaded gun”. Our obligation is to teach a “fire arms safety course.”
- Student’s can appreciate the importance of mathematics for solving problems within their discipline.

# Bioinformatics

- Rapidly emerging important new area in Biology
- Has radically changed the way Biological Research is being done
- Requires an interdisciplinary approach involving Biology, BioChemistry, Computer Science, and Mathematics
- More descriptive than analytical in its analysis
    - Sequence matching
    - Examining the 3-D structure of proteins

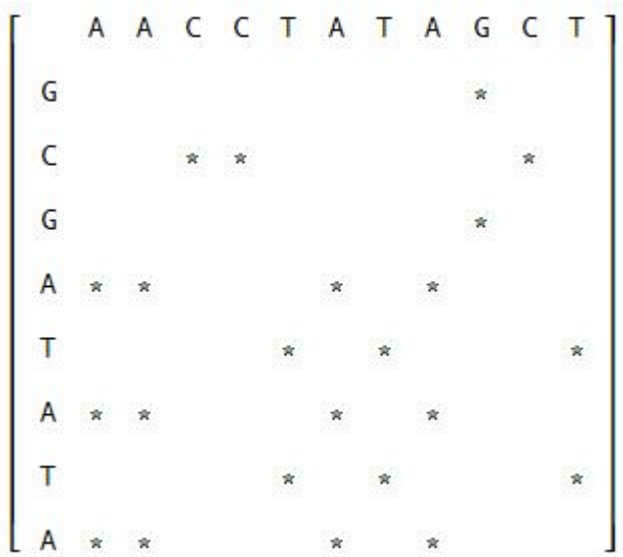
- We will only discuss the basics of sequence analysis for strings of DNA. We ignore the more complex (and interesting) protein sequences.

| Most Common Letters Used for DNA Nucleotide Sequences | | |
|---|---|---|
| *1-Letter Code* | *Nucleotide Name* | *Category (Comment)* |
| A | Adenine | Purine |
| C | Cytosine | Pyrimidine |
| G | Guanine | Purine |
| T | Thymine | Pyrimidine |
| N | Any nucleotide (any base) | (n/a) |
| R | A or G | Purine |
| Y | C or T | Pyrimidine |
| - | ——— | None (gap) |



# Dot Plots

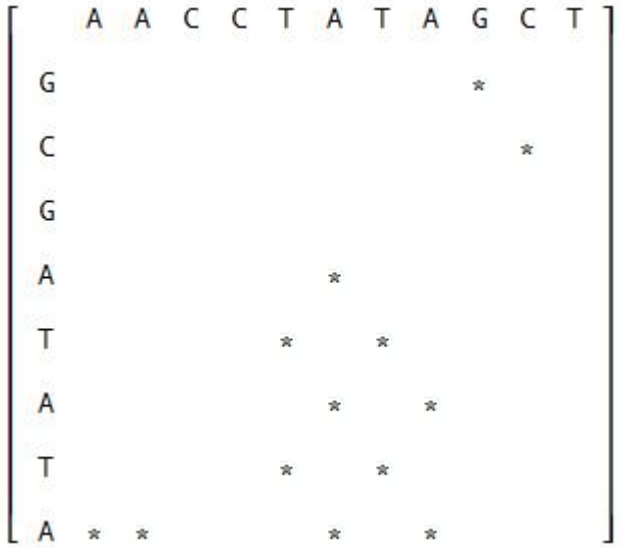
- This is the most elementary of the DNA String Comparison Tools.




Dot Plot for comparing two strings:

s1 := AACCTATAGCT
s2 := GCGATATA

- Result on previous slide is too simplistic
- Revise algorithm to only plot a star if at least one of the adjacent nucleotides matches the corresponding nucleotides of the other string




- Even with this revision this strategy will not work for comparing large sequences. Need another strategy.

- Problem of gaps:

  Perhaps one of these is a better allignment

  A A C C T A T A G C T

  G C G A T A T A - - -

  or

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

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

  Which is better?

- Scoring matrices - Several Schemes used based on different heuristics

  We use the following scheme for scoring matches and mismatches of nucleotides.

| | A | C | G | T |
|---|---|---|---|---|
| A | 1 | 0 | 0 | 0 |
| C | 0 | 1 | 0 | 0 |
| G | 0 | 0 | 1 | 0 |
| T | 0 | 0 | 0 | 1 |

Gaps (indels) are scored as -1

# Needleman – Wunsch Algorithm

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).

Place – (i – 1) in cell i of row 1 and – (j – 1) in cell j of column 1.

Starting in cell (i, i) with i $\geq$ 2, compute the following three values:

- The value in the adjacent cell to the left minus 1
- The value in the adjacent cell above minus 1
- 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.

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.

Repeat 2 a, b, and c and 3 above for all of the cells remaining in row i and column i.

Repeat steps 2, 3 and 4 until all of the cells of the matrix are evaluated.

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.

This is a dynamic programming algorithm. It starts filling the algorithm in the upper left cell and continues to the lower right cell.

CONTS

- Implementing the algorithm in Derive 6 Progrmming Language

s3 := ACTCG
s4 := ACAGTAG

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

Score = 2

- A good start, but not necessarily the best algorithm
- Consider the following situation

(NWAlign(ACGT, AAACACGTGTCT))$_3$

⌈ - - - - AC- - G - - T ⌉
⌊ AAACACGTGTCT ⌋

- First is clearly a subsequence of the second.
- Problem is that Needleman-Wunsch penalizes initial and terminal gaps the same as internal gaps.
- Internal gaps are clearly a result of indels whereas initial and terminal gaps are most likely a result of insufficient data collection
- Adjustment – the Semi Global Alignment Algorithm

  Scores initial and terminal gaps as 0 and internal gaps as -1 i.e. revises the initial step of Needleman-Wunsch

- Applying the Semi Global Alignment Algorithm to the same two sequences

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

This alignment makes much more sense. Note the score of 4. The Needleman-Wunsch Alignment had a score of -4.

# Summary

- "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 …"

National Research Council, BIO2010, Transforming Undergraduate Education for Future Research Biologists, National Academies Press, Washington, DC

- One emphasis of this course needs to be on constructing appropriate models and interpreting what the model can tell about the biological process
- Another needs to be on the construction of efficient descriptive algorithms to meet the needs of the new revolution in biological research.