ICTCM **Proceedings**

13TH ANNUAL

INTERNATIONAL CONFERENCE ON TECHNOLOGY IN COLLEGIATE MATHEMATICS

Addison Wesley

Boston San Francisco New York
London Toronto Sydney Tokyo Singapore Madrid
Mexico City Munich Paris Cape Town Hong Kong Montreal

in Proceedings of 13th ICTCM,
Nov. 2000 (Atlanta),
ISBN 0-201-761645
Pearson Education, Inc. 2002
246-249

INTRODUCTORY BIOLOGICAL SEQUENCE ANALYSIS THROUGH SPREADSHEETS

Stephen J. Merrill and Sandra E. Merrill
Marquette University
Department of Mathematics, Statistics and Computer Science
P.O. Box 1881
Milwaukee, WI 53201-1881
stevem@mscs.mu.edu sandra_merrill@yahoo.com

Abstract. Courses in mathematics for students of biology are often centered on calculus, even though the molecular revolution in biology has made the analysis of sequence data (DNA, RNA, and amino acids) the most important application of mathematical ideas to the science. Here we show the use of a spreadsheet to illustrate the techniques and provide a laboratory experience.

Introduction. Mathematical and statistical ideas, through modeling and analysis, have made numerous contributions to the biological and medical sciences. In recognition, most majors in biology require from 1-3 semesters of mathematics and statistics as part of their program. One important topic which is not currently introduced in these courses is **biological sequence analysis**, analysis of the strings of letters (coding for molecule names) resulting from techniques of molecular biology applied to DNA, RNA, and the amino acid sequence of proteins. Due to the discrete and nonnumeric character of this data, and the nature of the questions posed, standard techniques of calculus and statistics have little role to play in the analysis.

Spreadsheet software is commonly used in biological laboratories to assist in data acquisition, storage, analysis, and graphical presentation. It is also commonly present on student's personal computers and in college computer laboratories. This makes it an excellent tool to illustrate the mathematical concepts used in sequence analysis in a computationally relevant and flexible framework. In addition, use of spreadsheets can be quickly presented and make possible the use of large "real world" data sets. Although sequence analysis typically employs specialized software and interfaces which embed complex algorithms developed over many years, these programs conceal the mathematical concepts behind the computations and are thus easy to misuse and misunderstand. Using spreadsheets to introduce the basic definitions and explore their meanings should help to remove the "black box" feel of the research packages and provide a bridge to intelligent use of the growing software technology.

The use of spreadsheets in this context is not new. Robert F. Murphy at Carnegie Mellon University (http://www.bio.cmu.edu/Courses/03310) has developed a set of Microsoft Excel-based spreadsheets as part of an introductory course on Bioinformatics and the use of mathematical models in Biology. Here we extend that idea and use it to introduce this subject in a freshman or sophomore level course.

Meaningful Questions and Problems for sequence analysis. Most sequence analysis problems are variations of the following basic tasks:

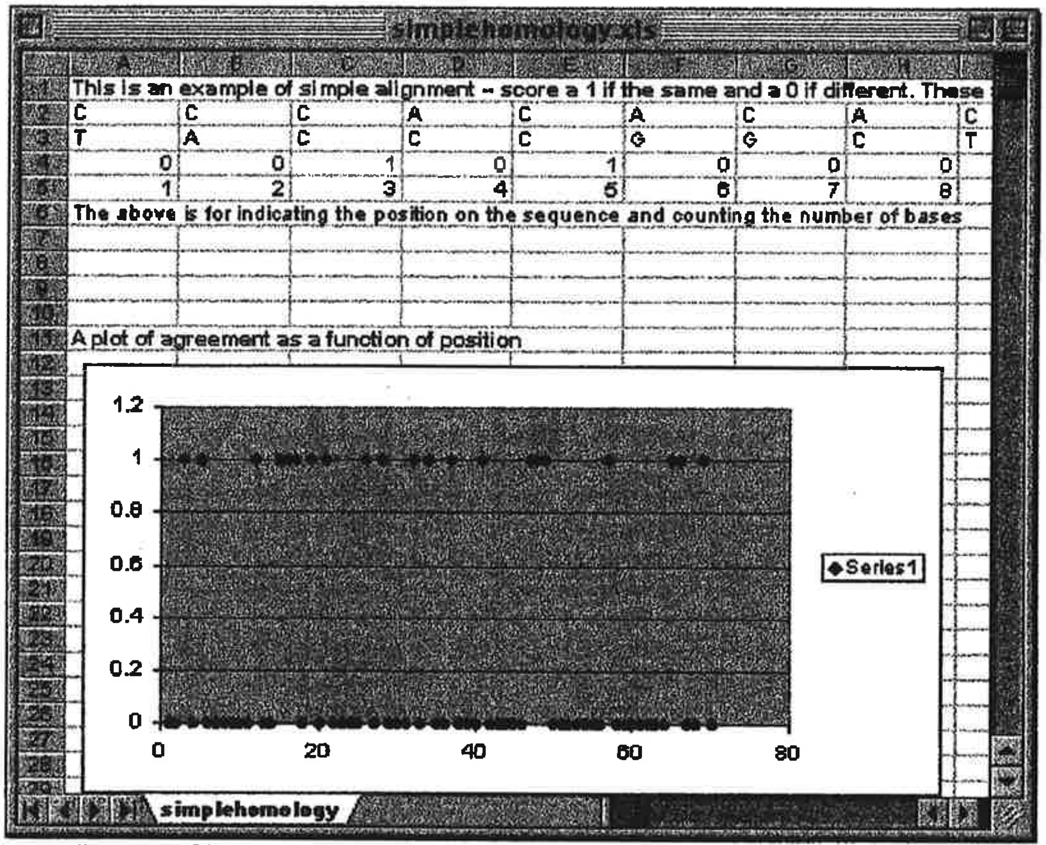
1) measuring the similarity between two strings,
2) finding instances of a particular pattern in a string,
3) describing the composition and properties of a string
4) graphing the evolutionary process and construction of phylogenetic trees to illustrate evolutionary distance between sequences

The strings currently most relevant to biology include DNA, RNA, and amino acids. The **primary structure** of these molecules is written as a list of repeating units. DNA is comprised of four such units: A (adenine), T (thymine), G(guanine), and C(cytosine). Regions of DNA, called genes, are transcribed into RNA (which also consists of a string of four letters), which are then translated into proteins. Proteins consist of a series of amino acids, twenty of which are most common. Information about the primary structure of known DNA, RNA, and protein sequences are stored in easily accessible databases. GenBank, for instance, currently has more than 9 million DNA sequences from 83,000 species stored (http://www.ncbi.nlm.nih.gov/Genbank). One protein database containing over 12,000 3-d structures is Brookhaven Protein Database (http://www.rcsb.org/pdb/).

Measuring the Similarity Between Two Strings. A fundamental question in sequence analysis is how similar is a given sequence to other sequences of interest. This question arises in several instances. For instance, suppose a gene has been identified and sequenced in humans. We would like to begin to investigate the function of the protein encoded by this gene. A typical starting place is to compare this sequence to a database of sequences found in other organisms. If close similarity is found between another gene and our unknown gene, we can begin to ask questions about the function of the similar gene. While not always the case, similar sequences sometimes indicate similarly shaped proteins and similar functions. Often we can find homologues, proteins with evolutionary ancestors to our protein.

Techniques and software for measuring similarity are abundant, but let's begin with a basic pair-wise comparison of two sequences. Comparing sequences and attributing a 'value' for how similar are two sequences is nicely illustrated using a spreadsheet by giving a point for every letter match. Spreadsheets easily demonstrate that sequences might have regions that are similar and regions that are not similar. Conserved regions, parts of a sequence that remain similar despite many years of evolution, often indicate areas of functional biological importance. For instance, the active site of a protein, where a molecule binds, is typically more conserved than other regions of a protein. In addition the chance of randomly finding a weak match to a sequence is demonstrated below.

Example 1 Simple homology. In this Microsoft Excel spread sheet, two DNA sequences of the same length from yeast are compared. The alignment score (simple homology of the sequences) is provided. Also random sequences are compared and a histogram of scores is displayed. Recalculating the sheet generates new sequences and comparisons.



	A	B	C	D	E	F	G	H	
1	This is an example of simple alignment -- score a 1 if the same and a 0 if different. These								
2	C	C	C	A	C	A	C	A	C
3	T	A	C	C	C	G	G	C	T
4	0	0	1	0	1	0	0	0	
5	1	2	3	4	5	6	7	8	
6	The above is for indicating the position on the sequence and counting the number of bases								
7									
8									
9									
10									
11	A plot of agreement as a function of position								

Figure 1. Simple homology worksheet

Finding Instances of a Particular Pattern in a String. Locating a pattern within a string is crucial for several aspects of sequence analysis. One important area includes finding the small percent of regions in DNA which correspond to genes and code for proteins. These regions are termed exons. Non-coding regions are called introns. Often there will be introns found within the gene which are removed (spliced) before its RNA is translated into a protein. Thus, for instance, even if we knew the sequence of all human DNA from the Human Genome Project, we still would not know how many human genes there are. However, we can recognize regions of the DNA sequence that contain patterns which we know from experience resemble genes. Gene-finding programs that utilize these patterns are plentiful. Typically they include looking for a three base pair sequence which starts every gene, called a start codon. Gene-finding software errs both on the side of finding regions which look like genes but are not (false positives) and of missing regions which are truly genes but don't appear to fit our selection criteria (false negatives). The exact criteria change as our knowledge of genes grows.

Other important pattern finding searches include looking for sites where restriction enzymes can recognize and cleave DNA, looking for splice sites where introns in an RNA sequence are removed, and searching for secondary structure motifs, patterns which correspond to a local folding arrangement of a protein sequence. Finding restriction sites is an interesting example both because of its importance in manipulating DNA and because of the hundreds of restriction enzymes currently available,

Describing the Composition and Properties of a String. Spreadsheets can also be used to display properties of a string. For instance, one might want to know the amount of G, C, A, and T's in a DNA sequence. This issue arises, for instance, in determining the point at which the two strands of DNA will separate, since G-C bonds are stronger than A-T bonds. Protein sequences are often analyzed for their amino acid content and isoelectric point, the pH at which the protein contains an equal number of negative and positive charges. Another useful analysis is to examine the protein sequence for hydrophobic (water repelling) regions, areas where the amino acids (aa) are uncharged and contain greasy sidechains. Spreadsheets can be used to draw hydropathy plots, which assign a hydrophobic score to each amino acid and then plot the score on the Y-axis and the position on the X-axis. This analysis quickly identifies hydrophobic regions that are unlikely to be exposed to solution and often represent interior regions of the protein or regions where the protein crosses a cellular membrane.

Example 2 Hydrophathy Plot. Using a 150 aa human protein, IL10, part of the system of cytokines in the immune system, this spreadsheet creates a hydropathy plot for this protein. As the structure of this protein is known, the relationship between this plot and the structure can be visualized.

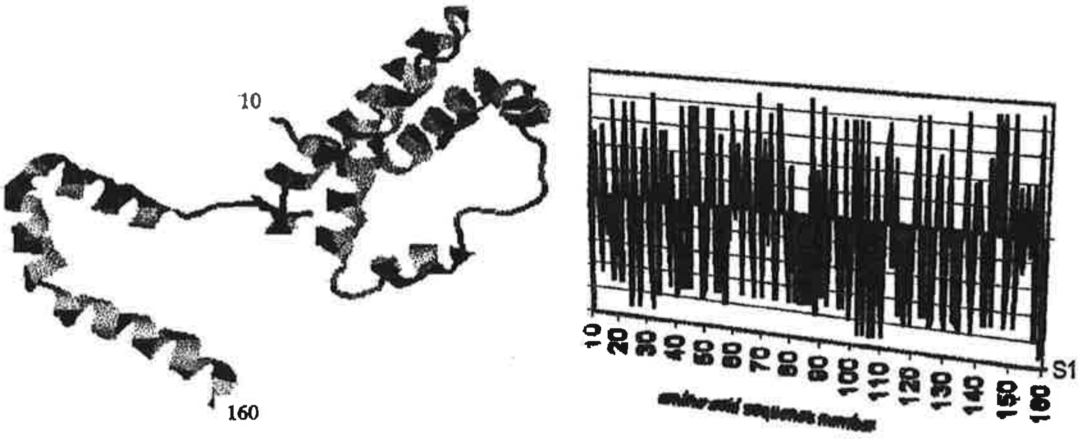


Figure 2. Structure of IL-10 (RasMol picture from Brookhaven PDB) and hydropathy plot for that protein. In the structure, positions that are hydrophobic (positive) are light and the dark color correspond to hydrophilic (negative) values in the plot.

Graphing Evolution and the Construction of Phylogenetic Trees. Evolutionary distance between sequences can be displayed using phylogenetic trees. Advanced software packages construct different types of trees using sophisticated algorithms and