EvolSeq 1.0

BioQUEST Curriculum Consortium

Summary
EvolSeq is an Excel workbook that simulates the molecular evolution of DNA sequences. The program begins with a single (random) sequence, then follows that sequence through time as it reproduces and mutates. Eventually, up to 20 evolutionary related sequences are generated. *EvolSeq* then calculates the genetic distances between each pair of DNA sequences, and also between the associated amino acid sequences.

Note: *EvolSeq* uses a macro to calculate genetic distances between sequences. Many antivirus software packages include options for disabling macros. These options must be turned off and macros must be enabled for *EvolSeq* to run on your computer.

Introduction
Phylogenetic trees are common and useful ways of graphically depicting patterns of evolution, both within and among individual evolutionary lineages. Among the many applications of phylogenetic analysis are:

- Identifying the exact source of ivory, wood, and other organic products suspected of being harvested illegally;
- Inferring migration and fragmentation events in a population's history; and
- Determining a gene's biochemical function by comparing it to other genes whose functions are known.

Behavioral, morphological, and molecular data can all be used to build trees, individually or in combination. Most current work utilizes molecular data sets, which will be the focus of this discussion.

For a given number of sequences, many distinct tree topologies (cladograms) are possible. This number increases combinatorially with the number of sequences (Table 1):

# of sequences	# of trees
3	3
4	15
5	105
6	945
7	10,395
8	135,135

10	2,027,025
15	2.13×10^{14}
20	8.20×10^{21}
25	1.19×10^{30}
30	4.95×10^{38}
N	$\frac{(2N-3)!}{2^{N-2}(N-2)!}$

Table 1. The number of distinct tree topologies for a given number of sequences. This table assumes that all trees are strictly bifurcating; i.e., that each divergence involves only two taxa. If multifurcations are allowed, the number of topologies rises still further.

A major issue in phylogenetics has been the development of computer algorithms to sort through this profusion of possible trees and determine which one(s) best explain a given set of sequence data. *EvolSeq* is designed to introduce students to this topic by generating data from which students can build phylogenetic trees using intuitive tree-building approaches. This approach gives students a solid foundation for approaching more sophisticated phylogenetic methods (see the **Resources** section below for references).

Imagine that we know the true evolutionary history of five related sequences; that is, we know the original sequence, the times at which pairs of sequences diverged from their common ancestor, and the timing and nature of individual mutations. Table 2 summarizes this information.

Initial Sequence:		**ATG CAA GGT TCA AGC**	
Time	**Event**	**Sequence(s)**	**Details (mutations only)**
1	Divergence	ABC vs. DE	
7	Mutation	ABC	11: C → T
19	Divergence	A vs. BC	
28	Mutation	A	3: G → C
36	Mutation	DE	7: G → A
51	Mutation	BC	13: A → G
67	Divergence	B vs. C	
72	Mutation	C	8: G → T
83	Mutation	C	1: A → C
86	Mutation	B	10: T → C
90	Divergence	D vs. E	
91	Mutation	D	4: C → A
92	Mutation	D	12: A → G
94	Mutation	E	5: A → G
96	Mutation	D	6: A → T
99	Mutation	E	14: G → A

Table 2. A summary table for the evolutionary history of six related sequences. Divergences represent the separation of a sequence into two distinct lineages, while mutations represent genetic changes within a single lineage. Mutation details are written in the form Position: Old → New, where "position" denotes the nucleotide position of the mutation, "old" denotes the nucleotide at that position prior to mutation, and "new" denotes the new (mutant) nucleotide.

We can depict these data graphically using any of the three commonly used classes of phylogenetic tree, depending on what aspects of the data we wish to highlight.

- A *cladogram* shows only the tree's topology (the pattern of evolutionary relationships among the sequences), discarding the temporal and mutational data. Lengths of individual branches in a cladogram thus have no real meaning.

- A *phylogram* shows both the tree's topology and the mutations occurring within each lineage, but does not include temporal data. Branchlengths in a phylogram thus correspond to the amount of evolutionary change on each branch.

- An *evolutionary tree* shows the tree's topology and the timing of each evolutionary divergence, but discards data on the amount of evolutionary change. Branchlengths in an evolutionary tree thus correspond to the duration of each branch.

We can draw a tree of each type corresponding to the data from Table 1:

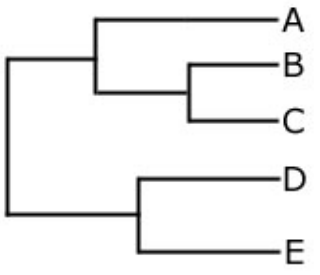

Fig. 1A. Cladogram.

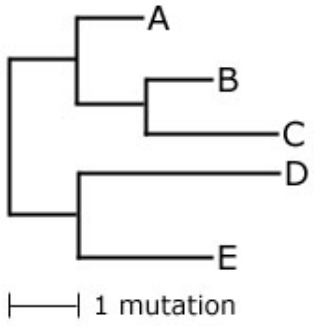

Fig. 1B. Phylogram.

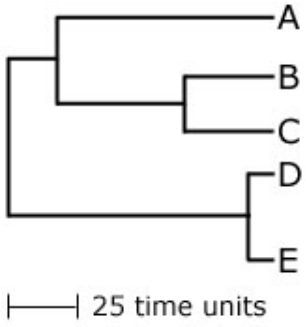

Fig. 1C. Evolutionary tree.

Two mathematical properties closely connected to the tree classes listed above are *additivity* and *ultrametricity*. A tree is said to be additive if the evolutionary distance between any two points in the tree is equal to the summed lengths of the branches connecting those points. It can be shown that all phylograms are additive, so long as no position within the sequence changes more than once.

A tree is said to be ultrametric if each tip of the tree is the same distance from the *root* (or base) of the tree. In other words, in an ultrametric tree, if the total length of branches connecting the root to one of the observed sequences is 15, then that is also the total length of branches connecting the base to each of the remaining observed sequences. Clearly, all evolutionary trees based on simultaneous samples are ultrametric.

Output: What do I see?

# of sequences:	5	(must be between 2 and 20 inclusive)			
Distance matrix: Nucleotide Sequences					
	Seq. 1	Seq. 2	Seq. 3	Seq. 4	Seq. 5
Seq. 1	0	4	4	6	8
Seq. 2		0	2	6	8
Seq. 3			0	6	8
Seq. 4				0	8
Seq. 5					0
Distance matrix: Amino Acid Sequences					
	Seq. 1	Seq. 2	Seq. 3	Seq. 4	Seq. 5
Seq. 1	0	3	2	4	5
Seq. 2		0	1	5	6
Seq. 3			0	4	5
Seq. 4				0	5
Seq. 5					0

Figure 2. Screen shot of *EvolSeq* , showing both the nucleotide and amino acid distance matrices. This example includes only five sequences; blank rows for up to 15 additional sequences have been hidden.

The main output of *EvolSeq* consists of two distance matrices. The first matrix, in blue, lists the number of nucleotide differences between each pair of sequences. The second matrix, in green, lists the number of amino acid differences between each pair.

The sequences themselves can also be viewed by selecting the "Sequences" tab at the bottom of the spreadsheet.

Controls: What can I do?

The only user-defined parameter in this version of *EvolSeq* is the number of sequences to be simulated. Enter this number in black-bordered cell B1 on the "Distance Matrices" sheet. The number of sequences must be an integer between 2 and 20, inclusive. As soon as you enter this parameter, the sheet runs its calculations and outputs the distance matrices.

If you want to run a new simulation without changing the number of sequences, use Excel's "Calculate Now" command. You can do this in the Calculation tab of the Preferences panel, or by using the keyboard shortcut (usually *Control* + "=" or *Command* + "=").

How it works: Model details

EvolSeq was designed primarily as a program for generating ultrametric and additive trees. As a result, the underlying evolutionary model makes a number of unrealistic simplifying assumptions. A more realistic model of molecular evolution is under development.

The simplest model for generating a random strictly bifurcating ultrametric tree with integer branchlengths is a discrete-time model in which a single branching event occurs at each time point and all branches between consecutive time intervals have length one:

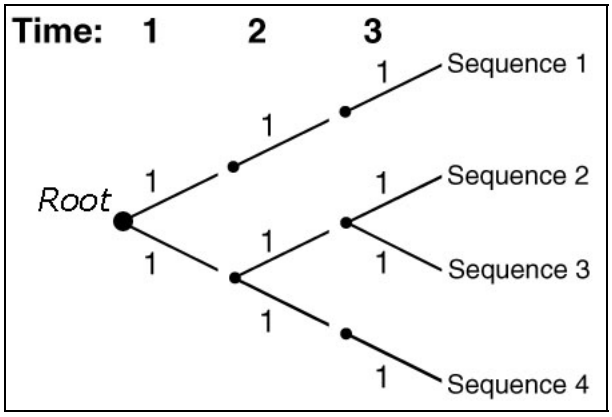

Figure 3. *EvolSeq*'s evolutionary model. At each time point, one randomly chosen sequence within the population diverges, then each sequence acquires one mutation. This model is biologically unrealistic but computationally simple.

This model requires at least $k = \frac{N(N-1)}{2} - 1$ variable positions within the sequence, where N is the final number of sequences desired. Moreover, because we want the amino acid sequences to yield an additive tree, we will permit only a single change within each codon (to avoid homoplasy). Therefore, we set the length of each nucleotide sequence to $3k$. The number of time intervals is equal to $N - 1$.

EvolSeq begins by computing these parameters from the user-entered value of N. All these calculations occur on the "Calculations" sheet. The program then assigns to each codon a random sequence of three digits corresponding to the four bases A, C, G, and T. Because each codon will mutate exactly once (under our assumed model), each codon is assigned a random rank that determines at what time point and in which sequence it will mutate, and is also assigned random values that determine what that mutation will be.

The program then runs through the computed number of time points. At each time point, a new sequence is generated from a randomly chosen existing sequence (the parent sequence), and then each sequence accumulates a single mutation. When the evolutionary process has been completed, *EvolSeq* randomly assigns new labels to each sequence. This is done so that sequences 1 and 2 will not always be the two oldest sequences, sequence 3 the next oldest, and so on.

The finished nucleotide sequences are then printed on the "Sequences" sheet. The program also translates the nucleotide sequences into amino acid sequences using the universal genetic code on the "Translation" worksheet, and prints these on the "Sequences" sheet as well. Finally, *EvolSeq* uses a custom macro to count the number of differences between each pair of nucleotide strings, and also between each pair of amino acid strings. These values are output as the distance matrices on the "Distance Matrices" sheet.

Ideas for classroom use

EvolSeq can be used to generate hands-on exercises in phylogenetic reconstruction. Working through this exercise helps students understand the basic methods of phylogenetic reconstruction and lays important groundwork both for studying more sophisticated methods (e.g. neighbor-joining, maximum parsimony, maximum likelihood) and for interpreting phylogenetic trees. For introductory purposes, we recommend using the nucleotide distance matrix and about 6 or 7 taxa. For more advanced work or longer-term projects, increase the number of taxa and/or use the amino acid distance matrix.

One approach is to have students work in groups of two or three for 5-10 minutes, then present their results and discuss their methods with the rest of the class. This exercise can follow an introductory mini-lecture on phylogenetic trees and their applications, including a discussion of evolutionary trees and phylograms and their respective relationships with the ultrametric and additive criteria.

Resources and further reading

Bioinformatics Web 2003. http://www.geocities.com/bioinformaticsweb/tutorial.html

Doyle J.J. and Gaut B.S. 2000. Evolution of genes and taxa: a primer. *Plant Molecular Biology* **42:** 1-23. Available online at http://bioinformatics.cs.vt.edu/~easychair/OrthologsParalogs/DoyleGaut_PlantMolBiol_2000.pdf

Golding B. and Morton D. 2003. Elementary Sequence Analysis, chapter 9. http://helix.biology.mcmaster.ca/chpt9.pdf

Guttman D.S. 2004. Bio472H1S Computational Genomics and Bioinformatics Course Notes, Feb. 4. http://www.botany.utoronto.ca/courses/bio472/Bio472-Phylogenetics1-4spp.pdf

Joyce D.E. 1996. Phylogeny and Reconstructing Phylogenetic Trees. http://aleph0.clarku.edu/~djoyce/java/Phyltree/cover.html

Mount D.W. 2001. Bioinformatics: Sequence and Genome Analysis, pp. 237-280. Cold Spring Harbor Laboratory Press: Cold Spring Harbor, New York.

Smith, C.M. 2004. CMS Molecular Biology Resource. http://restools.sdsc.edu/

Weston P.H. and Crisp M.D. 1998. Introduction to Phylogenetic Systematics. http://www.science.uts.edu.au/sasb/WestonCrisp.html

Terms and Conditions:

History:

Date: May 3, 2004
Title: *EvolSeq 1.0*
Name: Anton E. Weisstein and John R. Jungck
Institution: BioQUEST Curriculum Consortium, Beloit College
Acknowledgements: Support for this work was provided, in part, by the National Science Foundation Division of Undergraduate Education, the Howard Hughes Medical Institute, EOT-PACI, and the Shodor Foundation.
Modifications: None (original version).