MySSP

Version 1

Michael S. Rosenberg

Center for Evolutionary Functional Genomics
&
School of Life Sciences

Arizona State University

MySSP
Michael S. Rosenberg

Table of Contents

Citation

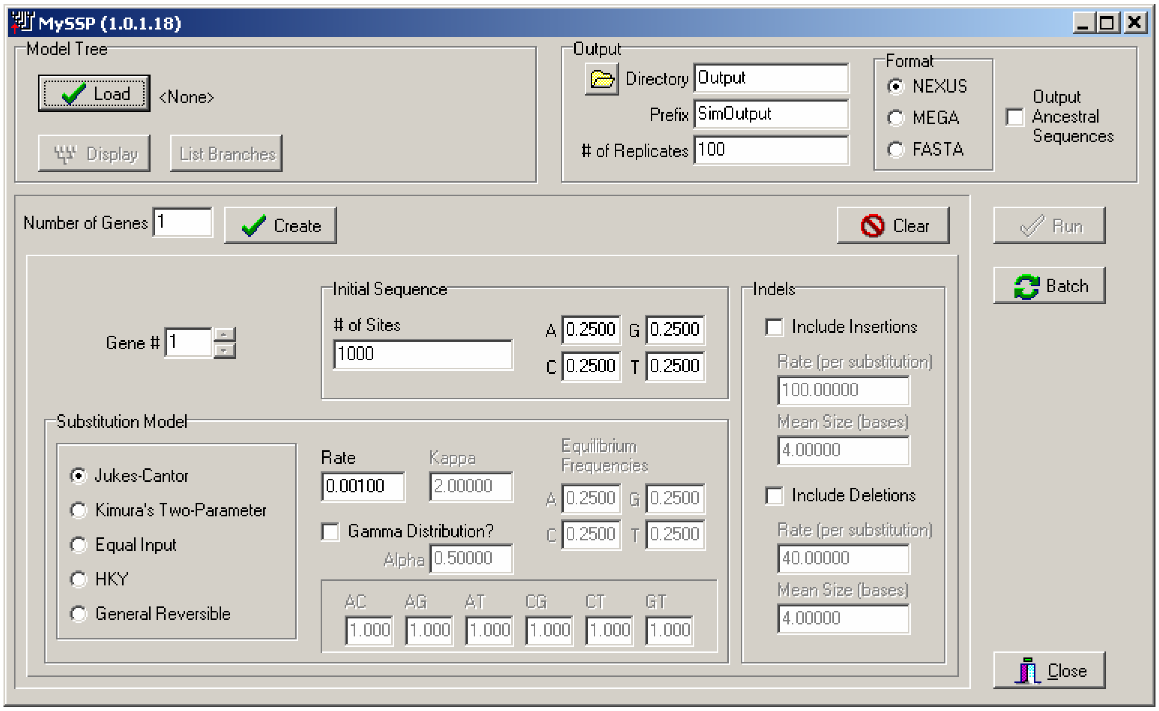
Rosenberg, M. S. 2005. *MySSP*: Non-stationary evolutionary sequence simulation, including indels. *Evolutionary Bioinformatics Online* 1:51–53.

This document was last altered on
April 12, 2006

Introduction

MySSP simulates sequence evolution across a fixed tree under a variety of models of nucleotide change. It has been gradually developed over a number of years and projects (Gadagkar *et al.* 2005; Rosenberg and Kumar 2001a, 2001b, 2003a; Rosenberg *et al.* 2003; Rosenberg and Kumar 2003b; Rosenberg 2005). It is designed to run under 32-bit Windows operating systems, although it also should work under common emulators.

The program runs in two modes: manual and batch. The batch mode can be used for running a variety of simulations simultaneously, as well as a means of allowing for non-stationary evolutionary processes. Both modes are executed from the GUI. The GUI consists of three parts: the model tree section, the output section, and the simulation parameter section.



<None>

Output
Directory Output
Prefix SimOutput
# of Replicates 100

Format
● NEXUS
○ MEGA
○ FASTA

□ Output Ancestral Sequences

Number of Genes 1

Gene # 1

Initial Sequence
# of Sites 1000
A 0.2500 G 0.2500
C 0.2500 T 0.2500

Indels
□ Include Insertions
Rate (per substitution) 100.00000
Mean Size (bases) 4.00000
□ Include Deletions
Rate (per substitution) 40.00000
Mean Size (bases) 4.00000

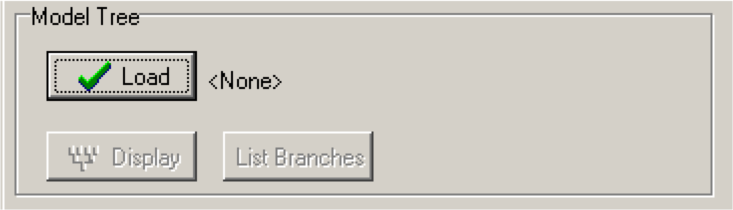
Substitution Model
● Jukes-Cantor
○ Kimura's Two-Parameter
○ Equal Input
○ HKY
○ General Reversible

Rate 0.00100
Kappa 2.00000
Equilibrium Frequencies
A 0.2500 G 0.2500
C 0.2500 T 0.2500
□ Gamma Distribution?
Alpha 0.50000

AC	AG	AT	CG	CT	GT
1.000	1.000	1.000	1.000	1.000	1.000

Tree Options

The model tree section is found in the upper left of the main window. This section is used to specify the model tree to use in the simulation.

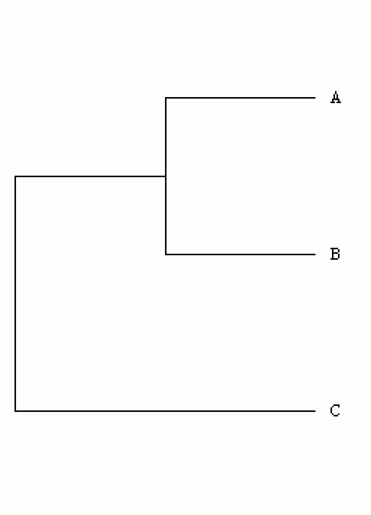


The LOAD button is used to specify a file containing the model tree. Model tree files should be simple text files containing the model in parenthetical format, *e.g.*,

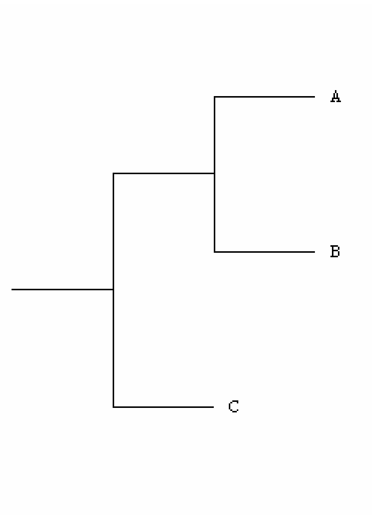
((A:0.05,B:0.05):0.05,C:0.1):0.0;

is a simple three-taxon tree. The tree must end with a semicolon. It is important to include branch lengths on the model tree, including the root (which should generally have a length of zero). By default, any unlabeled branch is given a length of 1.0, which can lead to unexpected results. The following shows how different input trees would be interpreted (each tree is scaled independently):

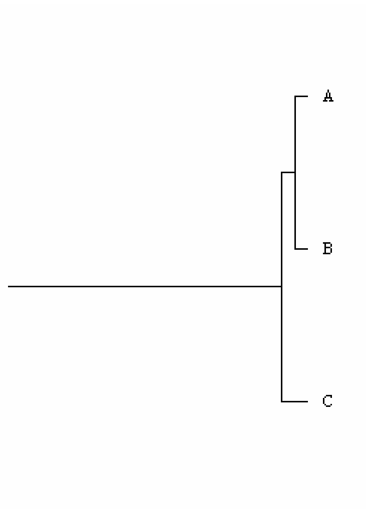
((A:0.05,B:0.05):0.05,C:0.1):0.0;



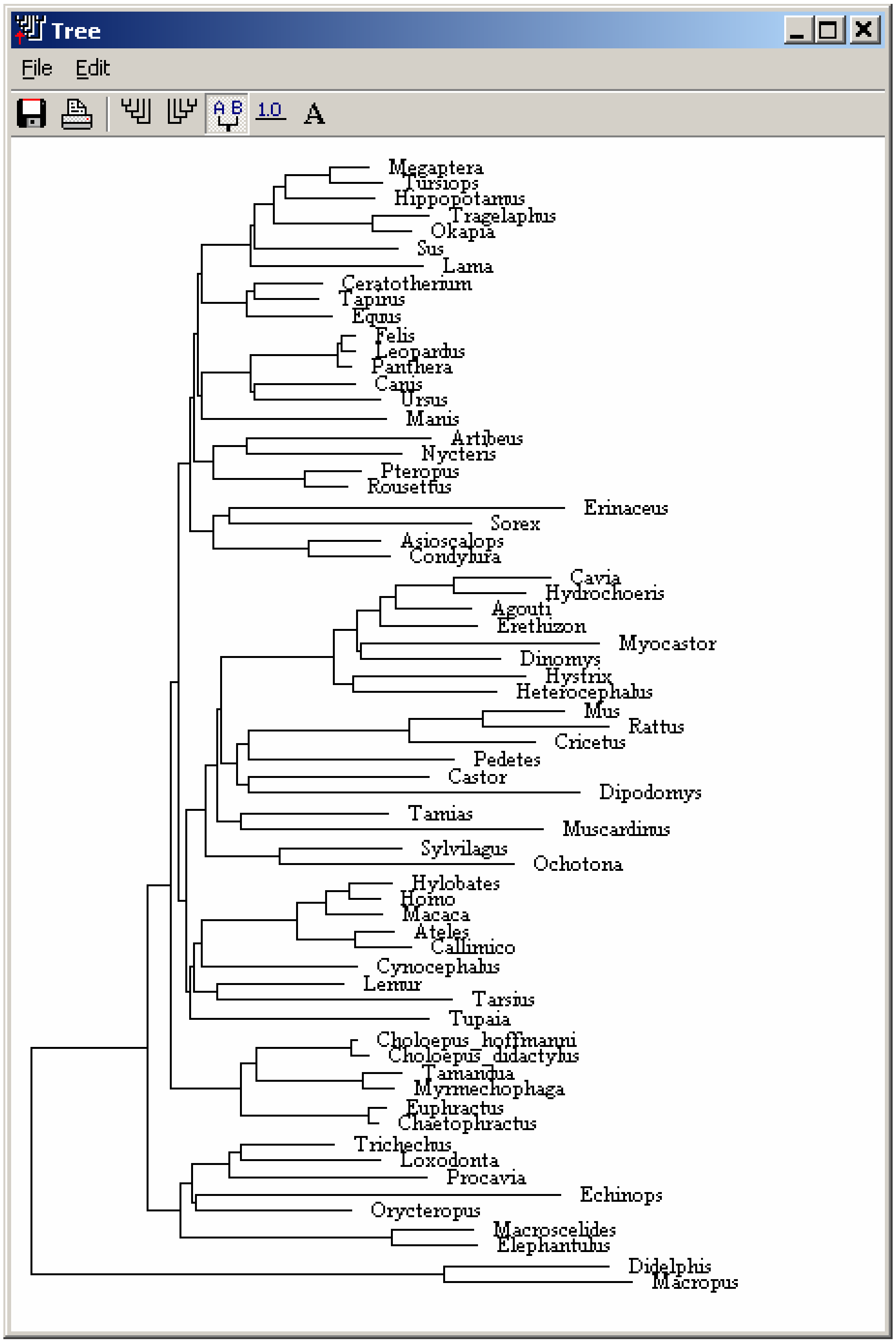
((A,B),C);



((A:0.05,B:0.05):0.05,C:0.1);



In general, *MySPP* assumes that all model trees are rooted. To display an input tree and confirm that it appears correct, press the DISPLAY button. This opens a window with a simple graphical representation of the tree. As an example, following is the 66-taxon tree from (Murphy *et al.* 2001).



The tree automatically scales to fit the window. The drawn tree can be saved to an image file, copied to the clipboard, or printed. The display options are fairly minimal (there are much better tree viewers available) and are primarily meant as confirmation that a tree file is being interpreted correctly; options include ladderizing the tree left or right and choosing whether to display taxa names and/or branch lengths.

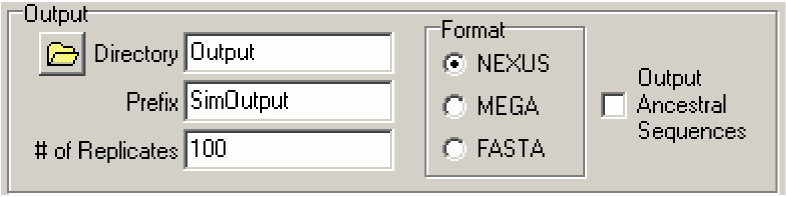
Pressing the final button in this section, LIST BRANCHES, will prompt for creation of a text file which will list every branch on the tree and the taxa which descend from it. Taking our ABC tree from above, the file would contain:

```
Branch#: All Descendent OTUs
0: A B C
1: A B
2: A
3: B
4: C
```

Each line contains the number of the branch and the complete list of descendent OTU's. By definition, the root of the tree is branch zero. These branch numbers can be used in conjunction with a batch script to specify model changes on specific parts of the tree (see batch scripts below).

Output Options

The output section is found in the upper right part of the main window. This section is used for specifying where and how data files will be output.



Directory: Output

Prefix: SimOutput

# of Replicates: 100

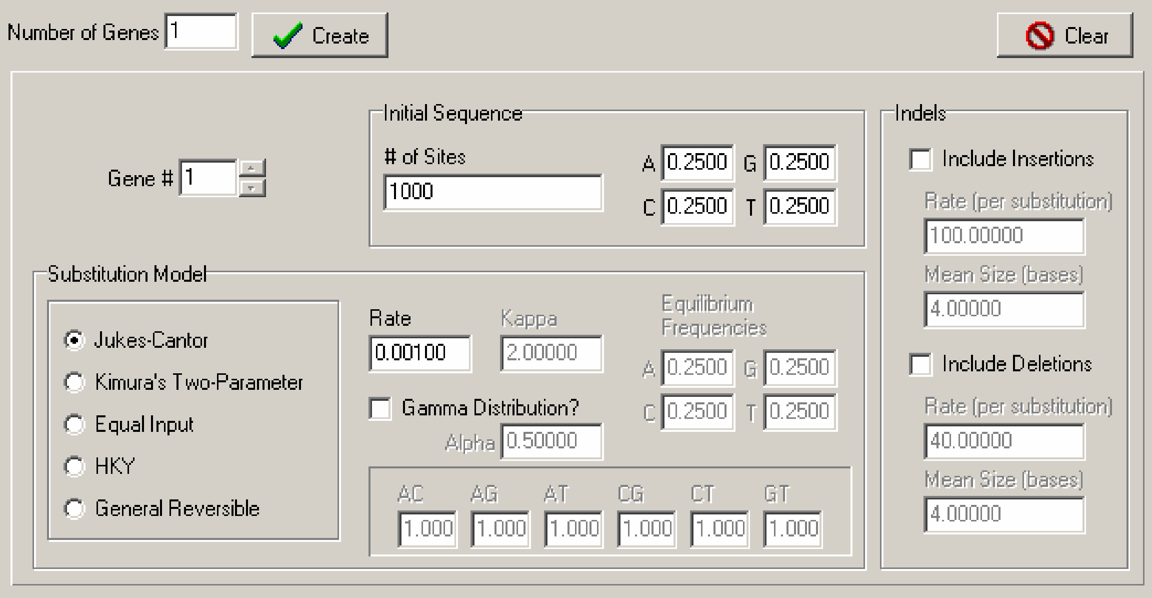
● NEXUS
○ MEGA
○ FASTA

□ Output Ancestral Sequences

There are five options. 1) Choose what directory/folder the output files will be generated in; when producing many files, it is often convenient to place the simulated data into a subfolder. 2) Choose a prefix for the output files. Each simulated data set will be written to a distinct file, named by the prefix with an appended number for each simulation replicate and an extension specified by the file format, *e.g.*, `SimOutput1.nex, SimOutput2.nex, SimOutput3.nex`, etc. **Note that *MySPP* will automatically write over files with identical names if they are found in the output directory.** 3) Choose the number of replicates to simulate, *i.e.*, how many data sets do you wish to generate with the specific parameter options. 4) Choose an output file type. *MySPP* can create output in three common sequence file formats: NEXUS files (.NEX), MEGA files (.MEG), and FASTA files (.FAS). 5) Choose whether to include ancestral sequences in the output file. Generally one is only interested in the tip (observed) sequences from the evolutionary process, but *MySPP* will optionally include all of the ancestral sequences (from the root upwards) in the data files. These sequences are placed as comments at the end of the data files. Each ancestral sequence is labeled by listing all of the terminal taxa names for which the sequence is an ancestor.

Model Options

The main bulk of the GUI allows choosing substitution models. The top of the window contains two buttons and an input box. The CLEAR button erases all choices and resets everything to the default. The NUMBER OF GENES box, allows the user to specify how many genes will be simulated simultaneously. Each gene can have its own set of independent parameters. Gene sequences are concatenated in the output, but specifically identified using character sets (or similar delimiters) in the NEXUS and MEGA output formats. By default only one gene is presumed to be simulated; to simulate more than one gene, change the number in the box and press the CREATE button. Pressing the CREATE button sets up the correct number of genes with the default model options; the number in the box has no meaning until after this button has been pressed.



Number of Genes 1

Gene # 1

Initial Sequence

# of Sites 1000

A 0.2500 G 0.2500
C 0.2500 T 0.2500

Substitution Model

● Jukes-Cantor
○ Kimura's Two-Parameter
○ Equal Input
○ HKY
○ General Reversible

Rate 0.00100

Kappa 2.00000

□ Gamma Distribution?

Alpha 0.50000

Equilibrium Frequencies
A 0.2500 G 0.2500
C 0.2500 T 0.2500

AC	AG	AT	CG	CT	GT
1.000	1.000	1.000	1.000	1.000	1.000

Indels

□ Include Insertions
Rate (per substitution) 100.00000
Mean Size (bases) 4.00000

□ Include Deletions
Rate (per substitution) 40.00000
Mean Size (bases) 4.00000

To set the parameters for each gene, first choose the gene with the GENE # box on the left side. Next, set the parameters describing the initial sequence (that is the ancestral sequence found at the root of the tree). This includes the sequence length and the starting nucleotide frequencies. When the simulation begins, an ancestral sequence is created by randomly choosing nucleotides with these expected proportions.

After choosing the initial parameters, one can set the substitution model which will be used to simulate the tree. There are five basic models to choose from: Jukes-Cantor (Jukes and Cantor 1969), Kimura's Two-Parameter (Kimura 1980), equal input, HKY (Hasegawa *et al.* 1985), and the general reversible model. These models are all explained in detail in other sources. Depending on the specific model, a variety of additional parameter options will become available:

Rate: Rate is available for all models. This value is used in combination with branch lengths to determine how many substitutions will occur on each branch. The expected number of substitutions per site is equal to rate × branch length. If the branches of the model already contain the exact desired rate, rate should be set to 1.0. This parameter can be used to easily scale genes to fast or slow (*e.g.*, by doubling or halving the branch lengths). In combination with non-stationarity (see below) it can be used to change the length of specific branches relative to the model tree.

Gamma Distribution: This can be used with all models. It models rate variation among sites, using the standard gamma model, with the shapr parameter specified by the ALPHA box.

Kappa: This is the ratio of transitions to transversions. It is used with Kimura's two-parameter model and the HKY model.

Equilibrium Frequencies: These are the expected frequencies of the nucleotides after evolution. Usually one will set these as identical to the initial frequencies, but they can be made to differ to allow for a general trend in changing frequency across the entire tree. These are used with the equal input, HKY, and general reversible models. To change nucleotide frequencies for part of the tree, you must change the model for the specific branches using the batch mode (see below).

AC/AG/AT/CG/CT/GT: These are the relative rates of the six types of reversible nucleotide substitutions used with the general reversible model. These are all relative values; the general use would be to set one of these to 1.0 and the others to values representing the relative frequency of the alternate pair wise substitution.

Insertion and deletions can also be set for the model. One can choose to model just insertions, just deletions, or both. For each indel type, one can set the rate of occurrence and the average size. Rates are specified relative to the substitution rate. Thus a value of 100 for insertion rate indicates that an insertion event occurs (on average) every 100 substitutions; thus larger numbers are less frequent. Actual numbers of indels are drawn from a Poisson distribution with the expected value determined by the rate and branch length. The size parameters indicate the mean size (drawn from a Poisson distribution truncated to not include zero) of an individual indel event, thus an event is not necessarily restricted to single bases as is often found in some indel models.

To include non-stationary substitution patterns and models, see the batch mode described below.

Running the Program

Once all of the parameters are set, the simulation can be performed by pressing the RUN button. As an alternate to the manual mode, *MySPP* can be run in a batch mode. Batch files are described below; to run a batch file, simply press the BATCH button and choose the file containing the commands.

Batch Commands

Running *MySPP* in batch mode allows one to run multiple simulations at once, allows one to use an external script generator to setup a variety of simulation conditions without requiring they be input by hand, and, most importantly, allows one to model non-stationary substitution patterns. Batch files are text files with a fairly simple structure. Lines starting with a "#" are considered comments and ignored. Commands are not case-sensitive. All batch commands end in a ";" and can thus be broken onto multiple lines.

One of the strengths of *MySPP* is that substitution models can be varied across the tree, by specifying different models for different branches. This is done as part of the GENE batch command (described below). When a model is set for a specific branch, it is set for that branch and all of its descendent branches. For a stationary process, all that needs to be done is to set a model for the root node (by definition, #0). To change the model for a subset of the tree, one simply needs to declare a new model for the branch at the base of the subtree where the change should occur. Branch numbers can be found by examining the output from the LIST BRANCHES button described above.\

Order is important. Each time a new model is declared, it is applied to the branch and all of its descendents; thus if you were to declare model A for a subtree and the declare model B for an ancestral branch of the subtree, model B would overwrite model A. In order to keep both, you must declare the ancestral models first followed by descendent models. As many (or as few) models can be declared as is necessary.

Note that all simulation parameters should be set with each model, not just those that are different from the ancestor. Each model is declared independently and by default represents a simple Jukes-Cantor model. One use of this flexibility could be to change the length of a single internal branch relative to the model tree, without actually having to modify the model tree. To do this: 1) set the model for the root node; 2) set the identical model for the target branch, with the only difference being the rate parameter; 3) set the original model for each of the descendent nodes of the target branch. Following this example, only the target branch will have the new rate.

The batch commands are as follows:

Tree

Equivalent to pressing the LOAD button, this is used to specify the file containing the model tree.

Syntax: TREE *treefile*;

Output

Sets the various output parameters described in the output section of the GUI.

Syntax: OUTPUT PREFIX=*prefix* DIR=*directory* NREPS=*number* FORMAT=NEX|MEG|FAS ANC=NO|YES;

PREFIX sets the output file prefix

DIR sets the output directory

NREPS sets the number of simulation replicates to perform

FORMAT sets the output file format. Valid values are NEX (for NEXUS), MEG (for MEGA), and FAS (for FASTA).

ANC controls whether ancestral sequences are included in the output. Valid values are YES and NO.

ClearGenes

Equivalent to pressing the CLEAR button, this resets all of the gene and substitution patterns to the default.

Syntax: CLEARGENES;

NGenes

Sets the number of genes to be simulated.

Syntax: NGENES *number*;

Gene

Sets all of the initial and substitution conditions for a specific to-be-simulated gene. Each gene (the number is set using the NGENES command) should be declared separately. This command is a bit more complex than the others, having an initial section which declares the gene and the starting parameters, and a model section which can be repeated multiple times if the model is changed on various modes.

Syntax:

```
Gene number NSites=number StartA=number StartC=number
    StartG=number StartT=number
    / branchnumber: Model=JC|K2|EqInp|HKY|GenRes Rate=number
        Kappa=number EqA=number EqC=number EqG=number EqT=number
        GenResPar=(number, number, number, number, number, number)
        GammaAlpha=number DoIns=Yes|No InsRate=number
        InsSize=number DoDel=Yes|No DelRate=number DelSize=number
    / branchnumber: Model=JC|K2|EqInp|HKY|GenRes Rate=number
        Kappa=number EqA=number EqC=number EqG=number EqT=number
        GenResPar=(number, number, number, number, number, number)
        GammaAlpha=number DoIns=Yes|No InsRate=number
        InsSize=number DoDel=Yes|No DelRate=number DelSize=number
    etc. (ending with ";")
```

The number immediate following the Gene command specifies the gene to which the remaining parameters apply.

NSites sets the sequence length of the initial sequence at the root of the tree

StartA, StartC, StartG, StartT sets expected frequencies of the nucleotides in the initial sequence at the root of the tree.

After these initial conditions are set, the substitution model should be specified. This is done by placing a "/" followed by a number indicating the node to which the model will apply. As described above, order is important. By definition, the root of the tree is node 0; other node numbers can be found by pressing the List Branches button. After the final model, use a semi-colon to end the command.

Model describes the basic substitution model. Valid values are JC, K2, EqInp, HKY, and GenRes.

Rate sets the branch length multiplier. Expected number of substations per site is equal to the branch length × rate.

Kappa sets the transition/transversion ratio. This is only used in the K2 and HKY models

EqA, EqC, EqG, EqT sets the expected equilibrium nucleotide frequencies. These are only used in the EqInp, HKY, and GenRes models.

GenResPar sets the parameters for the general reversible model. The six parameters are enclosed in parentheses and must appear in a specific order: (AC, AG, AT, CG, CT, GT).

GAMMAALPHA sets the shape parameter for substitution rate variation among sites. A value of zero indicates no rate variation.

DOINS sets whether insertion events should be modeled. Valid values are YES and NO.

INSRATE sets the rate of insertion events.

INSSIZE sets the mean size (number of nucleotides) of insertion events

DODEL sets whether insertion events should be modeled. Valid values are YES and NO.

DELRATE sets the rate of insertion events.

DELSIZE sets the mean size (number of nucleotides) of insertion events

Execute

Starts the simulation based on all of the previously specified parameters. Equivalent to pressing the RUN button.

Syntax: EXECUTE;

Quit

Closes *MySPP*.

Syntax: QUIT;

Examples

Following are some small sample batch files (found in the Examples subdirectory) demonstrating how to use *MySPP*.

Example 1

```
Tree 3_taxon_tree.txt;
NGenes 1;
Output Prefix=Example1_ Dir=ExampleOutput NReps=5 Format=nex;
Gene 1 NSites=100 StartA=0.2 StartC=0.3 StartG=0.3 StartT=0.2
  / 0: Model=HKY Rate=1.0 Kappa=3.2 GammaAlpha=0.0 EqA=0.2 EqC=0.3
       EqG=0.3 EqT=0.2;
Execute;
```

This simple example demonstrates the simulation of a single model for a single tree.

Example 2

```
Tree 3_taxon_tree.txt;
NGenes 2;
Output Prefix=Example2_ Dir=ExampleOutput NReps=5 Format=nex;
Gene 1 NSites=100 StartA=0.2 StartC=0.3 StartG=0.3 StartT=0.2
  / 0: Model=HKY Rate=1.0 Kappa=3.2 GammaAlpha=0.0 EqA=0.2 EqC=0.3
       EqG=0.3 EqT=0.2;
Gene 2 NSites=500 StartA=0.25 StartC=0.25 StartG=0.25 StartT=0.25
  / 0: Model=JC Rate=1.0 GammaAlpha=0.5 DoIns=yes InsRate=50 InsSize=6;
Execute;
```

This example demonstrates the simulation of 2 genes with different parameters simultaneously.

Example 3

```
Tree Murphy_mammal_tree.txt;
NGenes 1;
Output Prefix=Example3_ Dir=ExampleOutput NReps=5 Format=nex;
Gene 1 NSites=100 StartA=0.2 StartC=0.3 StartG=0.3 StartT=0.2
  / 0: Model=HKY Rate=1.0 Kappa=3.2 GammaAlpha=0.5 EqA=0.2 EqC=0.3
       EqG=0.3 EqT=0.2
  / 88: Model=HKY Rate=2.0 Kappa=3.2 GammaAlpha=0.5 EqA=0.15 EqC=0.35
        EqG=0.35 EqT=0.15;
Execute;
```

This examples demonstrates the simulation of a single gene for the 66-taxon tree, including a change of the model for a specific branch. Branch 88 is the branch at the root of the primate sub-tree and for the purposes of this example, the rate of evolution was doubled and the equilibrium frequencies altered.

Example 4

```
Tree 3_taxon_tree.txt;
NGenes 1;
Output Prefix=Example1_ Dir=ExampleOutput NReps=5 Format=nex;
Gene 1 NSites=100 StartA=0.2 StartC=0.3 StartG=0.3 StartT=0.2
  / 0: Model=HKY Rate=1.0 Kappa=3.2 GammaAlpha=0.0 EqA=0.2 EqC=0.3
       EqG=0.3 EqT=0.2;
Execute;

ClearGenes;

NGenes 2;
Output Prefix=Example2_ Dir=ExampleOutput NReps=5 Format=nex;
Gene 1 NSites=100 StartA=0.2 StartC=0.3 StartG=0.3 StartT=0.2
  / 0: Model=HKY Rate=1.0 Kappa=3.2 GammaAlpha=0.0 EqA=0.2 EqC=0.3
       EqG=0.3 EqT=0.2;
Gene 2 NSites=500 StartA=0.25 StartC=0.25 StartG=0.25 StartT=0.25
  / 0: Model=JC Rate=1.0 GammaAlpha=0.5 DoIns=yes InsRate=50 InsSize=6;
Execute;

ClearGenes;

Tree Murphy_mammal_tree.txt;
NGenes 1;
Output Prefix=Example3_ Dir=ExampleOutput NReps=5 Format=nex;
Gene 1 NSites=100 StartA=0.2 StartC=0.3 StartG=0.3 StartT=0.2
  / 0: Model=HKY Rate=1.0 Kappa=3.2 GammaAlpha=0.5 EqA=0.2 EqC=0.3
        EqG=0.3 EqT=0.2
  / 88: Model=HKY Rate=2.0 Kappa=3.2 GammaAlpha=0.5 EqA=0.15 EqC=0.35
        EqG=0.35 EqT=0.15;
Execute;

Quit;
```

This example includes all of the first three examples combined in a single batch file to show how multiple simulations can be conducted at once.

Technical Problems

For technical problems, please contact me at msr@asu.edu.

Release History

- 1.0.3.5 (April 12, 2006)
  - Fixed a bug involving gamma distributed rates that occasionally caused an unexpected math error, particularly when simulating branches with very large numbers of substitutions per site.
- 1.0.2.2 (August 2, 2005)
  - Fixed a bug due to changing gamma rates for different portions of the tree
  - Removed the limitation on using variable models which included both gamma rates and insertion events.
- 1.0.1.21 (July 22, 2005)
  - First public release of *MySSP*.

References

Gadagkar, S. R., M. S. Rosenberg, and S. Kumar. 2005. Inferring species phylogenies from multiple genes: Concatenated sequence tree versus consensus gene tree. *Journal of Experimental Zoology. B. Molecular and Developmental Evolution* 304B:64-74.

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* 22:160-174.

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pp. 21-132 in *Mammalian Protein Metabolism*, H. N. Munro, ed. New York: Academic Press.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base subsitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* 16:111-120.

Murphy, W. J., E. Eizirik, W. E. Johnson, Y. P. Zhang, O. A. Ryder, and S. J. O'Brien. 2001. Molecular phylogenetics and the origins of placental mammals. *Nature* 409:614-618.

Rosenberg, M. S. 2005. Evolutionary distance estimation and fidelity of pair wise sequence alignment. *BMC Bioinformatics* 6:102.

Rosenberg, M. S., and S. Kumar. 2001a. Incomplete taxon sampling is not a problem for phylogenetic inference. *Proceedings of the National Academy of Sciences USA* 98:10751-10756.

Rosenberg, M. S., and S. Kumar. 2001b. Traditional phylogenetic reconstruction methods reconstruct shallow and deep evolutionary relationships equally well. *Molecular Biology and Evolution* 18:1823-1827.

Rosenberg, M. S., and S. Kumar. 2003a. Heterogeneity of nucleotide frequencies among evolutionary lineages and phylogenetic inference. *Molecular Biology and Evolution* 20:610-621.

Rosenberg, M. S., and S. Kumar. 2003b. Taxon sampling, bioinformatics, and phylogenomics. *Systematic Biology* 52:119-124.

Rosenberg, M. S., S. Subramanian, and S. Kumar. 2003. Patterns of transitional mutation biases within and among mammalian genomes. *Molecular Biology and Evolution* 20:988-993.