SpedeSTEM 0.9.5 September 23rd, 2010

Bryan Carstens & Daniel Ence

SpedeSTEM is a program written in *java* that allows users to conduct a hierarchical analysis of species limits as described in Carstens & Dewey (2010). Essentially, SpedeSTEM allows its users to identify the best model of lineage composition using information theory for systems where individuals can be assigned to clusters. Ideally, these clusters will be identified using other sources of data such as morphological differentiation or geographic isolation, and in ideals case may represent formal taxonomic designations such as subspecies or races. SpedeSTEM incorporates STEM, a program that computes an analytical solution for the Maximum Likelihood Species Tree given some number of gene trees (Kubatko *et al*. 2009). As such, users are advised to read the accompanying guide to getting good results with STEM, as well as the paper that describes this program. There are several components to SpedeSTEM, including options for replicated subsampling, which is best used with the large data sets common to phylogeographic sampling.

Dependencies

- These instructions assume that the user has already installed PAUP* (SWOFFORD 2002), R, and the R-libraries *ape* (Paradis *et al*. 2004) and *gee*. Note that programs other than PAUP* can be used to estimate the gene trees, so long as they produce a midpoint-rooted ultrametric tree consistent with the molecular clock (read about gene tree requirements for STEM in its readme file).
- This program has been successfully run on Mac OS X 10.5.8 Mac OS X 10.6 and requires *java* 1.5 or greater and *R* 2.5 or greater.
- Version-numbers of *java* and *R* can be verified by typing "`java -version`" and "`R --version`" at the command-line. Note that the java version-command has one hyphen and the R version-command has two hyphens; that wasn't a typo.

Directions

1. Unzip the archive:

This will create several folders:

- *jar* contains the SpeDeSTEM.jar file
- *myotis_data* contains the example empirical dataset (Carstens& Dewey 2010).
- *related_manuscripts* contains several papers cited in this readme.
- *resources* contains two required R scripts.
- *simData* contains an example data set from Treatment I used in the program note that describes SpedeSTEM (Ence& Carstens in review). *simData* contains a set of nexus data files that can be used to verify proper function of the executable after installation.
- *STEMv1.1a* contains the STEM distribution package

The current version of SpeDeSTEM is 0.9.5.
The version for your .jar file can be verified by typing
"`java -Xmx512m -jar ./jar/SpeDeSTEM.jar --version`"

2. Execute SpedeSTEM with "`java -Xmx512m -jar ./jar/SpeDeSTEM.jar`" at the command line of a UNIX / LINUX terminal.

This begins a series of prompts for information (file paths, algorithm parameters, etc.) necessary to run SpeDeSTEM.

Command-line argument explanations:

`data`: This is the directory where the ".nex" files are located.

`grp`: This is the file that contains the list of species- and subspecies-memberships for each allele.

`mod`: This is the file that contains the sequence models and STEM multipliers for each loci.

`run`: This is the path for a directory that will be created to contain SpeDeSTEM results files.

`srch`: The type of search used by PAUP* to estimate genetrees. The default is NNI, other options include BANDB, TBR, SPR, NJ.

`stemdir`: This is the path to the directory that contains the STEM executable.

`stemex`: This is the name of the STEM exectuble file (probably STEM) to be used.

`rep`: This integer is the number of replicates to be run.

`theta`: The per locus value of theta, used by STEM to convert the subs/site branch lengths of the gene trees to the coalescent units of the species tree.

`auto`: This is a comma-delimited list of flags that directs which steps in the algorithm are to be executed. The list of currently accepted flags is as follows: sample, `paup`, `command`, `perm`, `stem`, `clean`, `quit`.

Explanation of auto flags:
`sample`: Directs SpeDeSTEM to subsample from the loaded files.

`paup` command: Directs SpeDeSTEM to execute PAUP* on the subsampled data files.

NOTE: The paup-command flag can be different on different systems. It is what the user would type to start PAUP* from the command-line (including any paths, etc.).
`Perm`: Directs SpeDeSTEM to test all possible permutations
of subspecies within a species in STEM.

`STEM`: Directs SpeDeSTEM to run STEM.

`Clean`: Directs SpeDeSTEM to delete all the intermediate results files,
leaving on the directory "STEMoutput"

`Quit`: Directs SpeDeSTEM to quit after successful completion of all replicates.

3. SpeDeSTEM can also be run in a batch mode, where all the parameters are supplied as command-line arguments by the user. The usage for batch mode is as follows:

`java -Xmx512m -jar ./jar/SpeDeSTEM.jar -data` <path to data files> `-grp` <path to the group file> `-mod` <path to SeqMod file> `-run` <path to a folder to be created for this run> `-srch` <NNI, SPR, TBR> `-stemdir` <path to STEM directory> `-stemex` <name of STEM executable> `-rep` <number of replicates> `-auto` <sample, paup*, perm, STEM, quit>

Items in red are specified by the user.

FILES NEEDED:
Data files should be in non-interleaved nexus format. Data from each locus should be contained in a separate nexus file, and the order of taxa should be consistent across loci.

The **Group file** is a tab-delimited file containing hierarchical information related to the membership of individuals to groups (subspecies, races, populations) and the membership of groups to species. In addition, the group file contains a column for sampling frequencies, defined as the number of alleles per replicate divided by the total number of alleles in that group for that locus. For example, if a group has 10 alleles, and you would like to subsample three alleles per replicate, then the value for each cell in the **Group file** should be 3/10=0.3. Note that theoretical research by Hird et al. (2010) suggests that 3 alleles per lineage is nearly as informative as 10 alleles per lineage for the purposes of estimating species phylogeny. Since the most computationally-intensive aspect of SpedeSTEM is the estimation of gene trees (by far), three alleles per lineage are adequate, particularly as a first-pass. Note that the user can conduct and analysis *without* subsampling by setting the sampling frequencies to "1", setting the `–rep` flag to 1, and calling on SpedeSTEM to sample all the alleles in the data set (i.e., set the sampling proportions in the **Group file** to 1.0).

The **SeqMod** file, also tab-delimited, contains information about the data file (column 1), the model of sequence evolution (column 2), and the STEM scaling factor (column 3). The model of sequence evolution is basically the "lset" line for a PAUP* analysis. This line must begin with "lset clock=no"; the search will be conduced without the clock, and then the trees will be rooted at the midpoint under the assumption of the molecular clock. When the subsampling option is used, we have assumed (see Hird et al. 2010) that model estimates from the full data set are appropriate to use for a subsampled data set. However, these assumptions have not been fully tested. Of course, parameters of sequence evolution can be estimated by PAUP* concomitantly with the tree search, albeit with a significant cost in the computation time. The STEM scaling factor is described in the STEM manual.

COMPUTATIONAL LIMITS OF SPEDESTEM.
Generating all hierarchical permutations is an important problem in the SpeDeSTEM work-flow. The problem can be stated as follows. Given a set of *n* lineages, with each lineage containing some number of *m* subgroups, generate all possible partitions of each subgroup, and generate all possible combinations of each partition of a given lineage with each partition of all the other lineages.

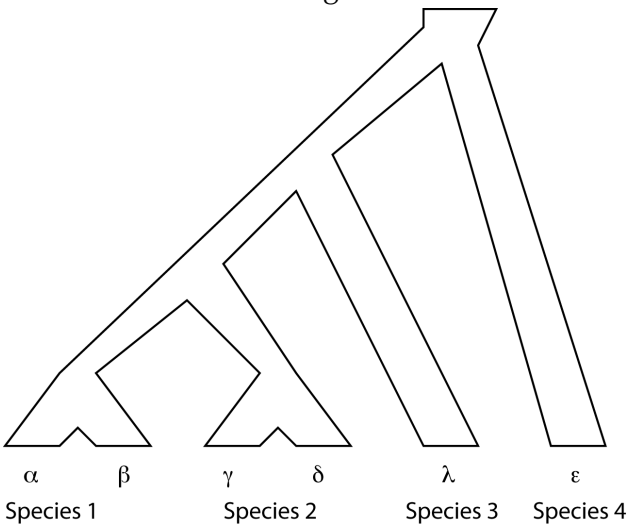
The number of possible partitions (ways of grouping subgroups within a lineage) of a group with *n* members is given by the *n*th Bell number. For example, given a group A, with subgroups a, b, and c, there are five possible partitions of A's subgroups. One could group them all separately, group them all together, or group any two of the subgroups together, leaving the third separate. The sequence of Bell numbers, beginning at zero, is 1,1,2,5,15,52,203,877,4140. Because this sequence grows so quickly and also because of computational memory issues, some practical limits were written into the program. SpeDeSTEM generates all possible partitions of up to eight subgroups, which gives 4140 possible partitions. If the user supplies a group file that include a lineage with more than eight subgroups, SpeDeSTEM will generate the first 4140 possible partitions, which are essentially randomly generated.

There is another multiplicative factor involved in combining a each partition for each lineage with all the other partitions of the other lineages. Therefore, the total number of permutations evaluated by one run of SpeDeSTEM, is total product of the Bell number's of the number of subgroups in each lineage. If a user gives a groups file that contains species A with five subgroups and species B with 4 subgroups, SpeDeSTEM will generate fifty-two partitions of A's subgroups and fifteen partitions of B's subgroups, and further will combine each of A's subgroups with each of B's subgroups, giving a total of 780 possible permutations.

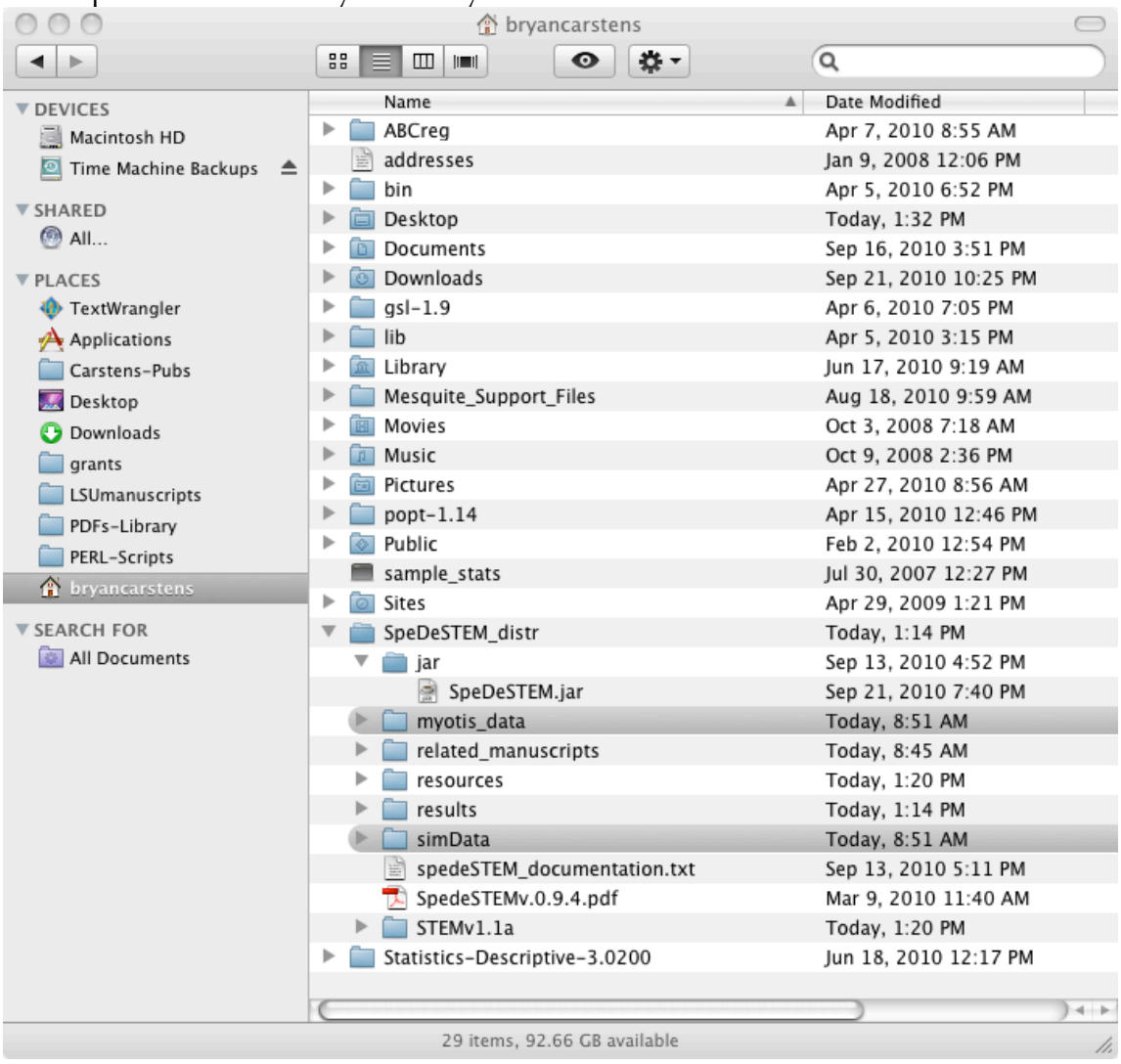
We have implemented SpedeSTEM with the goal of calculating hierarchical permutations in part because this allows us to increase the number of species-assignment models that can be considered. For users working in systems without subspecies or other *a priori* groupings, there is no reason why up to eight entities could not be treated as members of the same "species" and analyzed to determine the optimal assignment to lineages.

TEST DATA, EXAMPLE, RESULTS FILE:

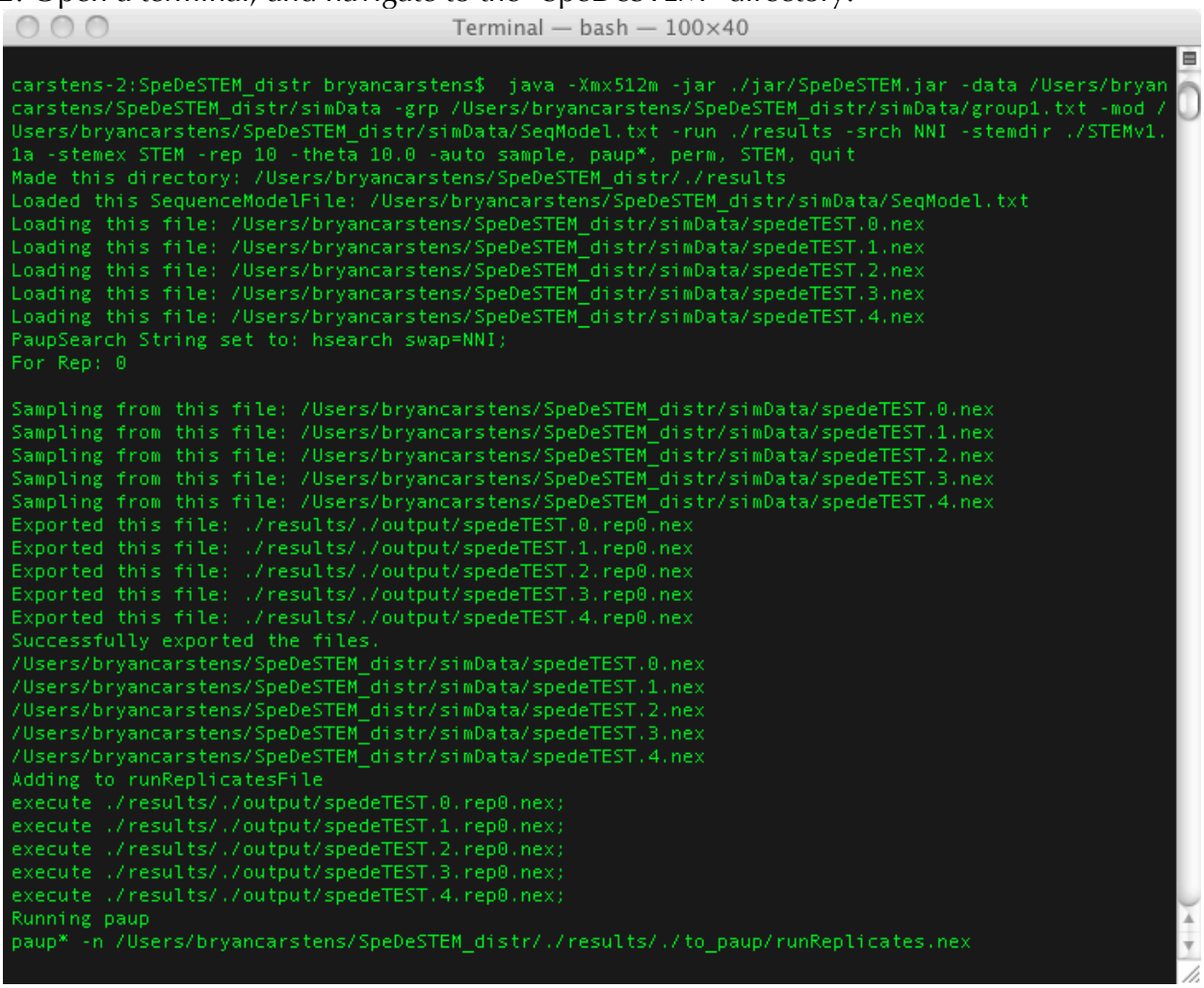
We include a simple data set for the purposes of demonstration. The data are sequence data simulated on genealogies simulated under a coalescent model corresponding to the one shown below. We generated these data for four species (number 1-4), with two of these species containing two putative populations or subspecies (α, β) or (γ, δ). Simulations were conducted using different combinations of branch length, and use SpedeSTEM to ask if (α, β) or (γ, δ) are genetically distinct lineages. For species 1 and 2, we simulate data for 100 individuals, and for species 3 and 4 we simulate data for 50 and 20 individuals. We also simulate sequence for an outgroup, not shown in the figure to the left. These sequences are described in the example group file "group1.txt".



Example:
1. unzip the archive in my directory:



2. Open a terminal, and navigate to the “SpeDeSTEM” directory:



3. Type the following at the command line (note that this command is stored in the simData folder as “Run_the_simData.txt”:

```
bryancarstens$  java -Xmx512m -jar ./jar/SpeDeSTEM.jar -data
/Users/bryancarstens/SpeDeSTEM_distr/simData -grp
/Users/bryancarstens/SpeDeSTEM_distr/simData/group1.txt -mod
/Users/bryancarstens/SpeDeSTEM_distr/simData/SeqModel.txt -run
./results -srch NNI -stemdir ./STEMv1.1a -stemex STEM -rep 10 -
theta 10.0 -auto sample, paup*, perm, STEM, quit
```

4. The screen out will report:
    a. that data from each locus has loaded
    b. that subsampling is occurring for each locus
    c. that replicated subsampling files are being exported
    d. that the tree searches in PAUP* are being conducted
    e. that trees from the replicate searches are being midpoint rooted using a molecular clock
    f. that STs are being calculated in STEM for each replicate with the taxonomic permutations

5. At this point, the analysis is complete. The results are located in the STEM_output folder. Shown are all species trees from each replicate, as well as a tab-delimited table reporting the results. I've pasted this below:

	replicate	. . .	**replicate**						
permutation	**0**		**9**	***ln*L (ave)**	**k**	**AIC (avg)**	**Δ_i**	**Model-likelihood**	***w*$_i$**
α_β, γ_δ, λ, ε, outgroup	-33.54	. . .	-32.08	-32.97	4	-24.968	0.000	1.000	0.695
α_β, γ, δ, λ, ε, outgroup	-31.54	. . .	-30.08	-31.25	4	-23.247	1.721	0.423	0.294
α, β, γ_δ, λ, ε, outgroup	-31.54	. . .	-30.09	-30.97	5	-15.969	8.999	0.011	0.008
α, β, γ, δ, λ, ε, outgroup	-29.54	. . .	-28.09	-29.25	5	-14.249	10.719	0.005	0.003

The left-most column depicts the permutations. Tests are conducted with both (α, β) and (γ, δ) place into the same OTU (1st row; membership denoted by the "_" character), with only (α, β) in the same OTU (2nd row), with only (γ, δ) in the same OTU (3rd row), and with all subspecies separated (4th row).

The likelihoods of the species tree for each permutation are shown for each replicate, as well as averaged across replicates. Also shown: the averaged AIC value, the AIC differences (Δ_i), model likelihoods, and model probabilities (w_i). Note that a detailed mathematical treatment of these values can be found in Anderson (2008), and a discussion of the interpretation of these values can be found in Carstens and Dewey (2010). Briefly, in an information theoretic perspective each of the permutations of lineage membership represents a hypothesis, and the AIC differences measure (in units that approximate Kullback-Leibler information) how much worse a particular model is than the best model (which is always listed in the first row). Since these units are abstract, a pair of transformations are conducted to convert these values to model probabilities (w_i), which essentially measure the proportion of the total model likelihood (i.e., across all models)

represented by that particular model. In this example, we have very strong support for the model that places (α, β) in the same lineage and (γ, δ) in a separate lineage. Note that these results are based on only ten permutations, likely far too few given our sampling. If we conduct 1000 permutations, the relative support for each model changes, and the rank order of the models could differ.

QUESTIONS:

Please send an e-mail to Dan Ence < dandence@gmail.com > and cc to Bryan Carstens < bryan.c.carstens@gmail.com >. One of us will get back to you as soon as we can.

REFERENCES:

Anderson DR (2008) *Model Based Inference in the Life Sciences* Springer, New York.

Carstens BC, Dewey TA (2010) Species Delimitation Using a Combined Coalescent and Information-Theoretic Approach: An Example from North American Myotis Bats. *Systematic Biology* **59**, 400-414.

Ence D, Carstens BC (in review) SpedeSTEM: A rapid and accurate method for species delimitation.

Kubatko LS, Carstens BC, Knowles LL (2009) STEM: Species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics*.

Paradis E, Claude J, Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289-290.

Swofford DL (2002) *PAUP*. Phylogenetic Analysis Using Parsimony (and other methods). Version 4*. Sinnauer Associates, Sunderland, MA.