Plan for simulations

Graham Jones

2011-10-11

1 Directory structure

This describes an example where there are 3 scenarios, A, B, C. Each scenario uses simulated data for 1 or 3 genes, 1 or 3 individuals per species, and two mutation rates, either 50e-9 or 100e-9. For each scenario there are 8 combinations leading to 3 files configA.txt, configB.txt, configC.txt, and 24 directories arranged like this:

```
scenarioA--configA.txt
           |-Ag1n1t50
           |-Ag1n3t50
           |-Ag3n1t50
           |-Ag3n3t50
           |-Ag1n1t100
           |-Ag1n3t100
           |-Ag3n1t100
           |-Ag3n3t100
scenarioB--configB.txt
           |-Bg1n1t50
           ...
scenarioC--configC.txt
           |-Cg1n1t50
           ...
```

The directory names consist of scenario letter, 'g' followed by number of genes, 'n' followed by number of individuals, 't' followed by mutation rate multiplied by 1e9, as an integer. This number (50 or 100 here) is the expected number of substitutions per billion sites per generation. Each of these directories has similar contents. For example `Ag3n3t50` contains the following files

Type	Files		
1MT	A1MTg3n3t50.txt		
2GT	A2GTg3n3t50r1-1.txt	A2GTg3n3t50r1-2.txt	A2GTg3n3t50r1-3.txt
	A2GTg3n3t50r2-1.txt	A2GTg3n3t50r2-2.txt	A2GTg3n3t50r2-3.txt
2SQ	A2SQg3n3t50r1-1.txt	A2SQg3n3t50r1-2.txt	A2SQg3n3t50r1-3.txt
	A2SQg3n3t50r2-1.txt	A2SQg3n3t50r2-2.txt	A2SQg3n3t50r2-3.txt
3BX	A3BXg3n3t50r1.xml		
	A3BXg3n3t50r2.xml		
4GTS	A4GTSg3n3t50r1-1.txt	A4GTSg3n3t50r1-2.txt	A4GTSg3n3t50r1-3.txt
	A4GTSg3n3t50r2-1.txt	A4GTSg3n3t50r2-2.txt	A4GTSg3n3t50r2-3.txt
4MTS	A4MTSg3n3t50r1.txt		
	A4MTSg3n3t50r2.txt		
4LOG	A4LOGg3n3t50r1.txt		
	A4LOGg3n3t50r2.txt		
5GTC	A5GTCg3n3t50r1-1.txt	A5GTCg3n3t50r1-2.txt	A5GTCg3n3t50r1-3.txt
	A5GTCg3n3t50r2-1.txt	A5GTCg3n3t50r2-2.txt	A5GTCg3n3t50r2-3.txt
5MTC	A5MTCg3n3t50r1.txt		
	A5MTCg3n3t50r2.txt		

The 'types', which also appear as part of the file name after the scenario letter 'A' are as follows.

- 1MT. The true multiply-labelled tree topology and node heights.
- 2GT. The simulated gene tree topology and node heights, one tree for each gene and each replicate.
- 2SQ. The simulated sequence alignments, one alignment for each gene and each replicate.
- 3BX. The BEAST XML files containing the sequence alignments, one per replicate.
- 4GTS. The gene trees sampled by BEAST. One set of gene trees for each gene and each replicate.
- 4MTS. The multiply-labelled trees sampled by BEAST. One set of trees for each replicate.
- 4LOG. The other parameters as sampled by BEAST. One sample for each replicate.
- 5GTC. The concensus gene trees produced by TreeAnnotator. One tree for each gene and each replicate.
- 5MTC. The concensus multiply-labelled trees produced by TreeAnnotator. One tree for each replicate.

The main point of the whole exercise is to compare the true multiply-labelled tree A1MTg3n3t50.txt with the estimated versions A5MTCg3n3t50r1.txt and A5MTCg3n3t50r2.txt

The file names all include 'g3n3t50' in order to make them identifiable without knowing the directory they are in. Not all of the files depend on all the configuration parameters. For example, the number of genes does not affect files with type 1MT or 5MTC.

2 Seeds for PRNGs

Seeds are required for each replicate, each simulated gene tree, and each simulated alignment. I used `http://www.random.org/integer-sets/` (100 sets, 100 numbers each, range 1-1e9), stored in a 2D array randomseeds[]. For gene i and replicate j, the seed used is randomseeds[i,j]. The same seed will be used for different scenarios, and different numbers of genes, individuals, mutation rates.