5S rRNA analysis : scripts to replicate analysis in Shim 2015

Heejung Shim

October 3, 2015

Contents

1 Installation of software packages

Please see BayesCAT Manual for installation of software packages relevant to our analysis including BayesCAT. BayesCAT Manual is available at

```
~/BayesCAT/doc/manual/
```

2 5S rRNA data

You can download data from BAli-Phy repository. Downloaded (and modified) 5S rRNA data is also provided in

```
~/BayesCAT/Shim_2015/data/5SrRNA.fasta
```

3 Analysis

If you followed the instruction in our github, you already have BayesCAT installed and have binary executable file BayesCAT in the directory ‘BayesCAT/src/’ or ‘BayesCAT/bin/’. We start with making a directory to perform an analysis.

```
cd ~/BayesCAT
mkdir Shim_2015/analysis/test
cd Shim_2015/analysis/test
```

3.1 Analysis of 5S rRNA

We can run **BayesCAT** on 5S rRNA data, using a script

```
/usr/bin/time -o time.txt -p ~/BayesCAT/bin/BayesCAT -seqfile ~/BayesCAT/
    Shim_2015/data/5SrRNA.fasta -seed 4 -iterations 1000000 -burnin 100000 -
    samplingIV 1000 -alpha_gamma 0.5 -alpha_kappa 0.5 -alpha_lambda 200 -
    alpha_A 13.3 -alpha_C 21.7 -alpha_G 23.1 -alpha_T 11.9 -alpha_r 100 -beta_r
     12200 -alpha_rd 3 -beta_rd 15
```

3.2 Summarize tree samples

‘REStree’ contains tree samples and we can summarize them using **summarize** for example

```
~/BayesCAT/bin/summarize REStree > tree.sum
```

Inside of tree.sum:

```
205 A1 ((((1,2),3),4),5)
170 A1 (((1,2),(3,4)),5)
131 A1 (((1,(2,3)),4),5)
109 A1 ((1,((2,3),4)),5)
205 0.205 0.205 A1
170 0.170 0.375 A2
```

```
131 0.131 0.506 A3
109 0.109 0.615 A4
```

3.3 Summarize multiple sequence alignments

‘RESAlignment’ contains multiple sequence alignment samples. First, we convert multiple sequence alignment samples to pairwise alignment samples using perl script ‘compute.Pair.Post.Prob.pl’ provided in ‘BayesCAT/scripts/perl/’, for example,

```
perl ~/BayesCAT/scripts/perl/compute.Pair.Post.Prob.pl RESAlignment alignment.
   sum
```

Then, ‘alignment.sum’ contains pairwise alignment samples. Now we can summarize/visualize multiple sequence alignment using the software package **FSA**. For example

```
cp ~/BayesCAT/Shim_2015/data/5SrRNA.fasta .
~/tmp/fsa-1.15.9/src/main/fsa --gui --load-probs alignment.sum 5SrRNA.fasta
```

Then, we can see **FSA** generates two files, ‘5SrRNA.fasta.probs’ and ‘5SrRNA.fasta.gui’. Now we can summarize/visualize alignment samples using for example

```
java -jar ~/tmp/fsa-1.15.9/display/mad.jar 5SrRNA.fasta
```

Please see FSA website for more options and detailed description of their output.

3.4 Summarize number of splits, edge length, and number of indel events

‘RESmnumIinAll’, ‘RESmnumDinAll’, and ‘RESmedgeLen’ contain number of insertion, number of deletion, and edge length for each (sampled) split. We can summarize them using a Script ‘get.summary.numID.edgelen.R’ provided in ‘/BayesCAT/scripts/R/’ for example

```
Rscript ~/BayesCAT/scripts/R/get.summary.numID.edgelen.R RESmnumIinAll
   RESmnumDinAll RESmedgeLen splits.sum 5
```

Here, 5 is number of sequences.

The output file ‘split.sum’ contains four columns as follows.

```
split numS.PP numID.PP edgelen.PP
10000 1 2.306 0.456168202
01000 1 3.54 0.464030069
00100 1 2.755 0.2641869629
00010 1 0.856 0.14698279726
00001 1 3.36 0.365654439
11000 0.414 0.161835748792271 0.112243413346135
10100 0.083 0.337349397590361 0.0464148006144578
10010 0.126 0.111111111111111 0.0804714332539682
10001 0.168 0.553571428571429 0.075931275297619
01100 0.305 0.380327868852459 0.111992973114754
01010 0.007 0 0.02506153
01001 0.161 0.372670807453416 0.0966510042608696
00110 0.303 0.0429042904290429 0.0904554959570957
00101 0.053 0.264150943396226 0.0442573445283019
00011 0.38 0.415789473684211 0.105108727868947
```

The first column is the split identifier. For example the identifier **11000** indicates the split which splits the first two sequences from the other three sequences. The second column contains a posterior probability for each split. The third and fourth columns contain posterior means of the number of indel events and the edge length given occurrence of each split.

Please see the description at the beginning of the script '/BayesCAT/scripts/R/get.summary.numID.edgelen.R' for detailed explanation for usage and options.

3.5 Summarize indel fragment sizes

'RESIleninAll' and 'RESDleninAll' contain samples of insertion and deletion fragment sizes. We can summarize them using a Script 'get.summary.fragmentSize.R' provided in '/BayesCAT/scripts/R/' for example

```
Rscript ~/BayesCAT/scripts/R/get.summary.fragmentSize.R RESIleninAll
   RESDleninAll indel.len.sum
```

The output file 'indel.len.sum' contains three rows, and each row contains a posterior estimate of realized indel (in the 1st row; realized insertion in the 2nd row; realized deletion in the 3rd row) fragment size distribution.

Please see the description at the beginning of the script '/BayesCAT/scripts/R/get.summary.fragmentSize.R' for detailed explanation of usage.

3.6 Summarize parameters

‘RESmGamma’, ‘RESmKappa’, ‘RESmP’, ‘RESmLambda’, ‘RESmMu’, ‘RESmR’, ‘RESmRi’, and ‘RESmRd’ contain samples of parameters, γ, κ, π, λ, μ, r, r_i, and r_d. We can summarize them using a Script ‘get.summary.param.R’ provided in ‘/BayesCAT/scripts/R/’ for example

```
Rscript ~/BayesCAT/scripts/R/get.summary.param.R RESmGamma gamma.sum 95 1
Rscript ~/BayesCAT/scripts/R/get.summary.param.R RESmP pi.sum 95 4
```

The script takes three arguments (input file, creditable interval, and number of parameters in the input file). The first output file ‘gamma.sum’ contains four columns: mean, median, and 95% CI for γ as follows:

```
3.589903 3.40015 1.540043 6.607065
```

The second output file ‘pi.sum’ contains four rows: each of rows contains mean, median, and 95% CI for each of π as follows:

```
0.173097312 0.172911 0.164125825 0.1852387
0.324600251 0.3225895 0.306639825 0.348690725
0.336307983 0.339181 0.3121102 0.34709065
0.165994433 0.1626005 0.150785525 0.186221475
```

Please see the description at the beginning of the script ‘/BayesCAT/scripts/R/get.summary.param.R’ for detailed explanation of usage.

3.7 Summarize other quantities

‘RESmnumD’, ‘RESmnumI’, and ‘RESmtotalEdgeLen’ contain samples for number of deletion, number of insertion, and total sum of branch lengths in a tree. We can summarize (mean, median, CI) them using a script ‘get.summary.param.R’.