

Parse MrBayes Ancestral States

- Fri 05 Jan 2018
- Johan.Nylander@{nrm|nbis}.se

The software MrBayes have the capacity of inferring ancestral character states. The results are summarized in the `.pstat` file. This script parses the MrBayes output and saves the actual states or sequence, with the highest posterior probability, as a fasta-formatted file.

Set up MrBayes to do the ancestral reconstructions

Example dataset with combined DNA and morphological data. When reading the data, the morphological data are automatically assigned to a separate partition (number 2). We will define one specific node where ancestral states should be estimated, then tell MrBayes to report the ancestral states for all characters in partition 2. The `Lset coding=variable` command is for accounting for coding bias in the morphological data. See the MrBayes manual for explanation. Also note that on your data, you might want to change default settings on substitution models, MCMC, etc.

#NEXUS

```
Begin MrBayes;
  Set autoclose=yes nowarn=yes;
  Execute mix.nex;
  Taxset MyTaxset = Andricus_curvator Andricus_kollari;
  Constraint MyConstraint = MyTaxset;
  Prset topologypr = constraints(MyConstraint);
  Lset applyto=(2) coding=variable;
  Report applyto=(2) ancstates=yes;
  Mcmc nruns=1 nchains=1 ngen=50000 checkpoint=no;
  Sump;
  Quit;
End;
```

Parse the output from MrBayes

Run the script with the `.pstat` file as input. The resulting sequence is written to standard out, while the details about state probabilities is printed on standard error:

```
$ ./parsemb.pl data/mix.nex.pstat
# MyConstraint
485 0=6.361512e-01 2=2.980698e-01 1=6.577902e-02
486 1=9.579908e-01 0=4.200920e-02
487 2=9.909044e-01 0=8.459564e-03 1=6.360484e-04
488 1=9.693519e-01 0=2.806066e-02 2=2.587463e-03
489 2=9.919600e-01 0=7.198033e-03 1=8.419819e-04
490 0=5.094045e-01 2=4.638707e-01 1=2.672478e-02
>MyConstraint Ancestral states at node MyConstraint
012120
```

Alternatively, if you wish to have the table (tab separated) and fasta in separate files:

```
$ ./parsemb.pl data/mix.nex.pstat > ancestral.fas 2> ancestral.txt
```

Example 2: Ancestral sequence estimation

Here we will report the ancestral sequence for a specific node in the tree. We will use an example file with DNA data.

MrBayes commands:

```
#NEXUS
Begin MrBayes;
  Set autoclose=yes nowarn=yes;
  Execute dna.nex;
  Taxset MyTaxset = Andricus_curvator Andricus_kollari;
  Constraint MyConstraint = MyTaxset;
  Prset topologypr = constraints(MyConstraint);
  Report ancstates=yes;
  Mcmc nruns=1 nchains=1 ngen=50000 checkpoint=no;
  Sump;
  Quit;
End;
```

Run the script:

Here we redirect the table to /dev/null, and fold the long sequence using the unix command fold:

```
$ ./parsemb.pl data/dna.nex.pstat 2> /dev/null | fold
>MyConstraint Ancestral states at node MyConstraint
TCTTGGTCCATTCTTGTGTGAGATATACGCTTTACTTGGTTCTCTATTTGGATGTGGCTCCATTTGGACAATGTGTATGA
TTGCTTTTGGATAGGTACAATGTAATAGTGAAAGGTTTGGCTGGGAAGCCCTTAACAATCACCGGTGCAATTATACGCATA
ATTGGCCTTTGGGTCTGGGCCATTATTTGGACTATTGCGCCAATGTTTGGATGGAATCGGTTTTATGTACCTGAAGGTAA
CATGACAGCTTGCAGGAAGTATTATTAAGTAAAGACTGGTTCTCGAGGTCTTACATCCTTGTATACAGTATCTTCGTAT
ACTATATGCCGCTTTTCCTTATCATATACAGTTACTATTTTATCATCTCAGCTGTATCTGCTCACGAAAAAGCAATGCGC
GAACAGGCCAAAAAGATGAACGTAGCTTCTCTACGTTTCATCTGACAATGCAACACAAGTGCTGAGCATAACTCGCAAA
GGTA
```

Alternatively, we may filter the ancestral sequence and say, for example, that we should only give the states where the pp is above 0.99. The other states are given as an “unknown” character (default is “?”, here we will use N):

```
$ ./parsemb.pl --cutoff 0.99 --unknown N data/dna.nex.pstat 2> /dev/null | fold
>MyConstraint Ancestral states at node MyConstraint
NNNNGGTCCATTCTTGTGTGAGATATACGCTTTACTTGGTTCTCTATTTGGATGTGGCTCCATTTGGACAATGTGTATGA
TTGCTTTTGGATAGGTACAATGTAATAGTGAAAGGTTTGGCTGGGAAGCCCTTAACAATCACCGGTGCAATTATACGCATA
ATTGGCCTTTGGGTCTGGGCCATTATTTGGACTATTGCGCCAATGTTTGGATGGAATCGGNNNTATGTACCTGAAGGTAA
CATGACAGCTTGCAGGAAGTATTATTAAGTAAAGACTGGTTCTCGAGGTCTTACATCCTTGTATACAGTATCTTCGTAT
ACTATATGCCGCTTTTCCTTATCATATACAGTTACTATTTTATCATCTCAGCTGTATCTGCTCACGAAAAAGCAATGCGC
GAACAGGCCAAAAAGATGAACGTAGCTTCTCTACGTTTCATCTGACAATGCAACACAAGTGCTGAGCATAACTCGCAAA
GGTA
```

Thanks to

Thanks to Dr. Robin Beck, Univ. Salford, for feature suggestions.