BUCKy
Bayesian Concordance Analysis

Workshop: Estimating Species Trees

Bret Larget

Cecile Ane

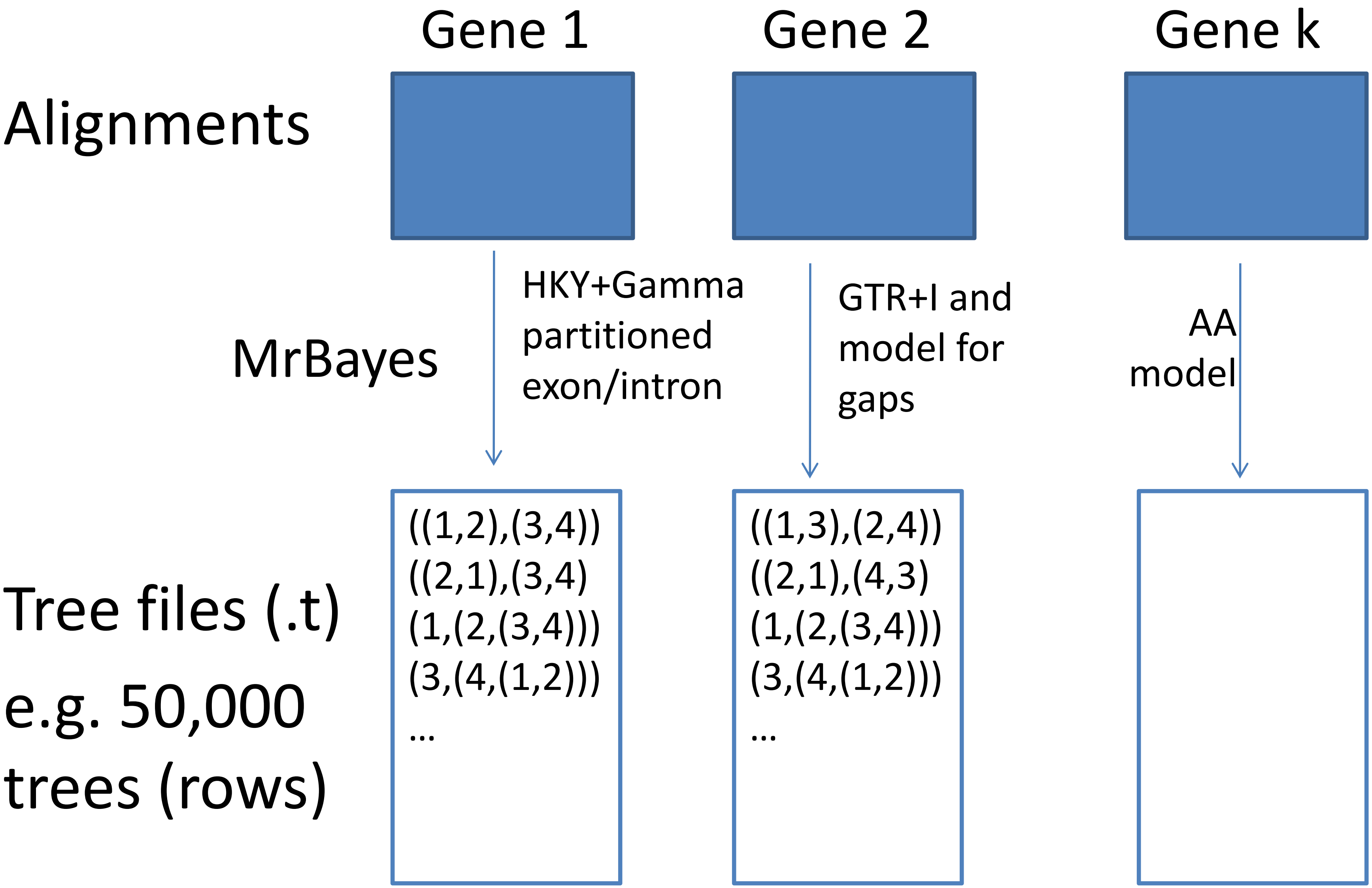
Step 0: Example data set

- Wild potatoes (Rodriguez *et al*. , in review)
- Go here: www.stat.wisc.edu/~ane/bucky/ and download the ‘example’ (.tgz)
- Let’s look at the data…
- Mac: re-install the program:

```
cd  ~/bin/source/bucky-1.3a/src/
make clean
make
mv bucky     ~/bin
mv mbsum  ~/bin
```

- Extract from archive, put ‘workshopEST’ into
    - On Mac: ~/bin/   (for instance)
    - On Windows: Desktop\ESTIMATING…\BUCKy…\

Step 1: MrBayes, individual loci



Gene 1

Gene 2

Gene k

Alignments

MrBayes

HKY+Gamma
partitioned
exon/intron

GTR+I and
model for
gaps

AA
model

Tree files (.t)

e.g. 50,000
trees (rows)

((1,2),(3,4))
((2,1),(3,4)
(1,(2,(3,4)))
(3,(4,(1,2)))
…

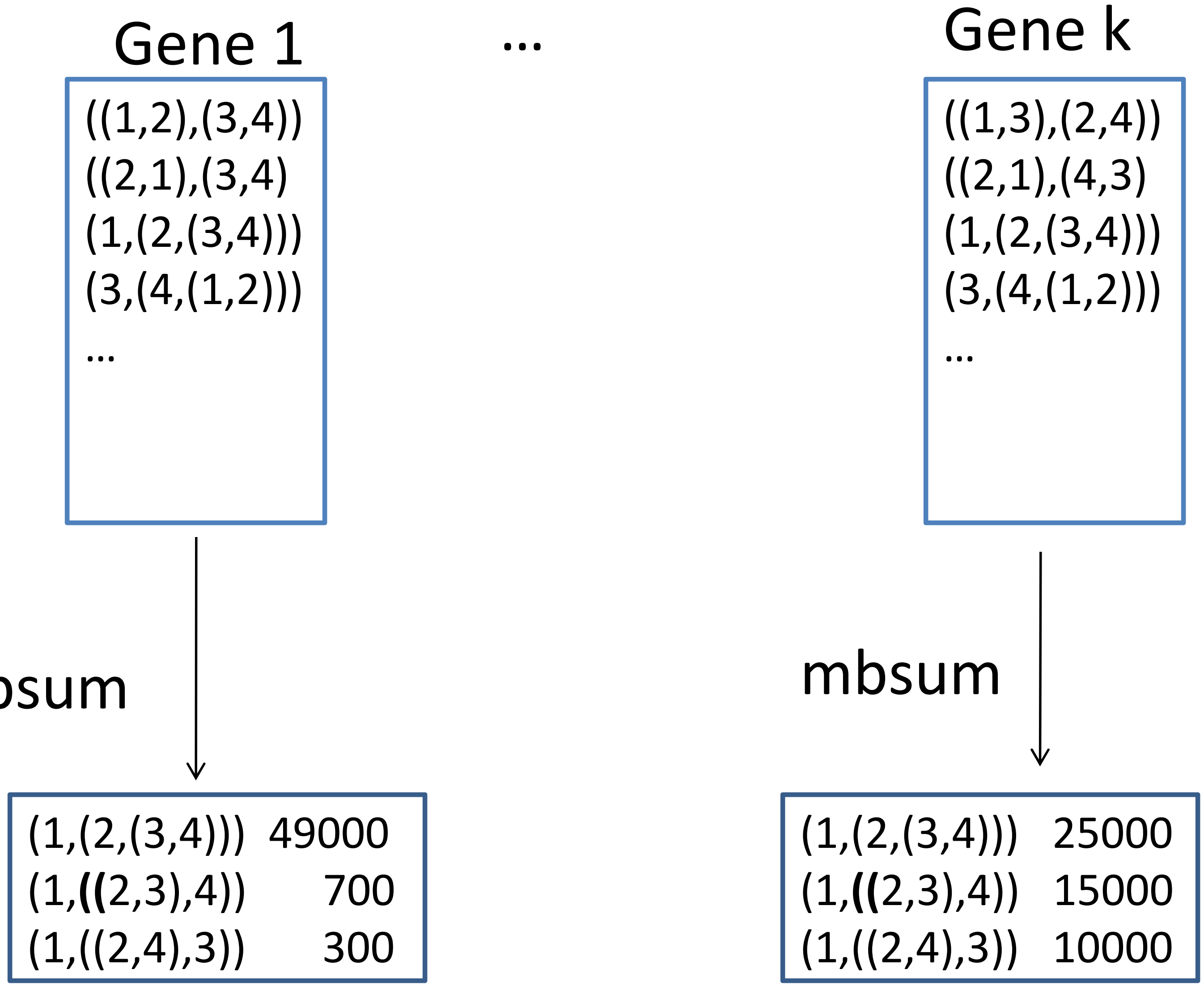
((1,3),(2,4))
((2,1),(4,3)
(1,(2,(3,4)))
(3,(4,(1,2)))
…

Step 1: MrBayes, individual loci

- Results from MrBayes on wild potatoes are in the file downloaded: workshopEST.tgz

- Extract from archive, put into
  - On Mac: ~/bin/ (for instance)
  - On Windows: Desktop\ESTIMATING…\BUCKy…\

Summarize MrBayes output: mbsum



Tree files (.t)

e.g. 50,000 trees (rows)

Gene 1

((1,2),(3,4))
((2,1),(3,4)
(1,(2,(3,4)))
(3,(4,(1,2)))
…

…

Gene k

((1,3),(2,4))
((2,1),(4,3)
(1,(2,(3,4)))
(3,(4,(1,2)))
…

mbsum

mbsum

Tree posterior prob. files

e.g. 200 rows

(1,(2,(3,4)))	49000
(1,**((**2,3),4))	700
(1,((2,4),3))	300

(1,(2,(3,4)))	25000
(1,**((**2,3),4))	15000
(1,((2,4),3))	10000

Summarize MrBayes output: mbsum

- Open workshopEST /workshop_commands (with Wordpad is okay)
- Mac: from a terminal window:

```
cd     ~/bin
./mbsum –h
```

Then use the commands in the file.

- Windows: from the start menu: run… -> cmd cd Desktop\ESTIMATING…\BUCKy…\

  to resize the window: right click -> properties -> layout -> Window size: 140

```
mbsum   -h
```

Summarize MrBayes output: mbsum

Options for mbsum:

-n # number of trees to skip, for burnin

-o name of output file

list of .t files, one per run

Summarize MrBayes output: mbsum

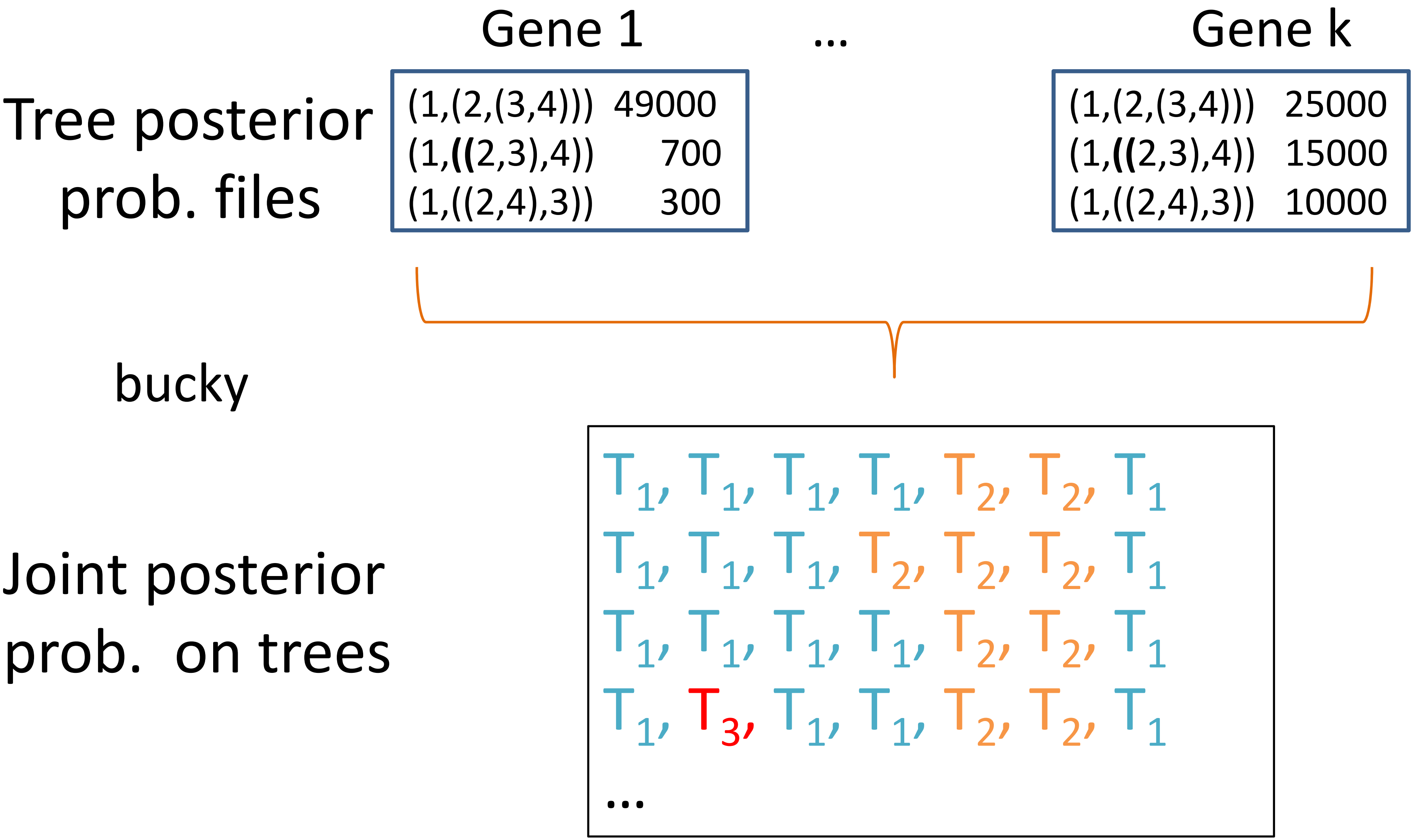
```
./mbsum –h
```

```
./mbsum -n 1001    -o    cos3.in   mrbayes_output/Potato_cos3.run?.t
./mbsum -n 1001    -o    cos5.in   mrbayes_output/Potato_cos5.run?.t
./mbsum -n 1001    -o    cos8.in   mrbayes_output/Potato_cos8.run?.t
./mbsum -n 1001    -o    cos9.in   mrbayes_output/Potato_cos9.run?.t
./mbsum -n 1001   -o   cos7br.in   mrbayes_output/Potato_cos7br.run?.t
./mbsum -n 1001    -o    cos9b.in   mrbayes_output/Potato_cos9b.run?.t
```

…

10b, 11, 1c, 3c, 5c, 2c

Step 2: bucky



Tree posterior prob. files

Gene 1	
(1,(2,(3,4)))	49000
(1,**((**2,3),4))	700
(1,((2,4),3))	300

…

Gene k	
(1,(2,(3,4)))	25000
(1,**((**2,3),4))	15000
(1,((2,4),3))	10000

bucky

Joint posterior prob. on trees

$T_1, T_1, T_1, T_1, T_2, T_2, T_1$

$T_1, T_1, T_1, T_2, T_2, T_2, T_1$

$T_1, T_1, T_1, T_1, T_2, T_2, T_1$

$T_1, T_3, T_1, T_1, T_2, T_2, T_1$

…

Step 2: bucky

Joint posterior prob. on trees

↓

$T_1, T_1, T_1, T_1, T_2, T_2, T_1$
$T_1, T_1, T_1, T_2, T_2, T_2, T_1$
$T_1, T_1, T_1, T_1, T_2, T_2, T_1$
$T_1, T_3, T_1, T_1, T_2, T_2, T_1$
…

Summaries:

- Number of **clusters**: 1 cluster = 1 group of genes that share the same topology
- Number of genes with a given clade = **CF** of the clade
  - Sample-wide
  - Genome-wide

Step 2: options for bucky: bucky -h

-n # generations. 10% more will be done for burnin.

-k # of independent runs

-a alpha   or --use-independence-prior

-c # chains (one cold, other heated)

-m #: heated chains use a=am, am^2, am^3 , etc.

--calculate-pairs   --create-sample-file

-o output file names

-s1 # -s2 #   to fix the seeds

-i name of file containing the list of input files

Step 2: bucky

```
./bucky -k 4 -a 0.1 -n 10000 -o a01 --calculate-pairs -i listInputfiles

./bucky -k 4 -a 0.1 -n 10000 -o a01 --calculate-pairs cos3 cos5 cos8 cos9 cos7br cos9b cos10b cos11 cos1c cos3c cos5c cos2c

./bucky -k 4 -a 0.1 -n 10000 -o a01 --calculate-pairs cos*.in
```

Step 2: bucky output

.cluster: posterior distribution of the # of clusters

.concordance: Gives

- estimated concordance tree, with sample-wide concordance factors as branch lengths and as # of genes supporting a clade.
- Concordance factors of major clades, with credibility intervals. For sample-wide and genome-wide CF.
- More info on CFs, with SD across runs.

.pairs: posterior probability that pairs of genes have the same tree.