BUCKy

Bayesian Untangling of Concordance Knots
(applied to yeast and other organisms)

Version 1.4.0, last updated 24 June 2010

Departments of Statistics and of Botany
University of Wisconsin - Madison
Medical Sciences Center,
1300 University Ave. Madison, WI 53706, USA.

Introduction

BUCKy is a program to analyze a multi-locus data sets with Bayesian Concordance Analysis (BCA), as described in Ané *et al.* (2007) and Larget *et al.* (2010). See Baum (2007) for the concepts of concordance factors and concordance trees. BCA accounts for biological processes like hybridization, incomplete lineage sorting or lateral gene transfer, which may result in different loci to have different genealogies. With BCA, each locus is assumed to have a unique genealogy, and different loci having different genealogies. The *a priori* level of discordance among loci is controlled by one parameter α.

BCA works in two steps: First, each locus is to be analyzed separately, with MrBayes for instance. Second, all these separate analyses are brought together to inform each other. BUCKy will perform this second step. BUCKy comes into two separate programs: `mbsum` and `bucky`. The first program `mbsum` summarizes the output produced by MrBayes from the analysis of an individual locus. The latter, `bucky`, takes the summaries produced by `mbsum` and performs the second step of BCA. These two programs were kept separate because `mbsum` is typically run just once, while `bucky` might be run several times independently, with or without the same parameters, with a subset of taxa or a subset of genes, etc.

Installation and Compilation

BUCKy is a command-line controlled program written is C++. It should be easily compiled and run on any Linux system or Mac OSX. Executable files compiled under various platforms are available at `www.stat.wisc.edu/~ane/bucky/` .

Installation from source code. Pick a directory where you want the BUCKy code to be. This directory will be called `$BUCKY_HOME` in this documentation. Download the `bucky-1.4.4.tgz` file and put it in `$BUCKY_HOME`. To open the compressed tar file with the BUCKy source code and example data, do these commands:

```
cd $BUCKY_HOME
tar -xzvf bucky-1.4.4.tgz
```

This creates a directory named `bucky` with subdirectories `bucky/data` and `bucky/src`.

Compilation. If you have gcc installed, compile the software with these commands.

```
cd $BUCKY_HOME/bucky/src
make
```

This will compile programs `mbsum` and `bucky`. It is suggested that copies of `mbsum` and `bucky` be put in `~/bin` if this directory is in your path. If you do not have gcc installed and the executable provided is not working on your system, you need to find the installer for gcc. On a Macintosh (version 10.3.9 or before), it may be in Applications/Installers/Developer Tools.

Running mbsum

Type this for a brief help message

```
mbsum --help
mbsum -h
```

Purpose and Output. It is advised to have one directory containing the MrBayes output of all individual locus analyses. Typically, in this directory each file of the form `*.t` is a MrBayes output file from one single locus. Use `mbsum` to summarize all files from the same locus. You want `mbsum` to create a file `<filename>.in` for each locus. The extension `.in` just means input (for later analysis by `bucky`). Output files `*.in` from `mbsum` will typically look like the following, containing a list of tree topologies and a tally representing the trees' posterior probabilities from a given locus (as obtained in the first step of BCA).

```
translate
       1 Scer,
       2 Spar,
       3 Smik,
       4 Skud,
       5 Sbay,
       6 Scas,
       7 Sklu,
       8 Calb;
(1,(2,(3,(4,(5,((6,7),8))))); 24239
(1,(2,(3,(4,(5,(6,(7,8)))))); 15000
(1,(2,(3,(4,(5,((6,8),7))))); 2983
(1,(2,(3,((4,5),((6,7),8)))); 2590
(1,(2,((3,((6,7),8)),(4,5)))); 2537
(1,(2,((3,(6,(7,8))),(4,5)))); 1097
(1,(2,(3,((4,5),(6,(7,8))))); 995
(1,(2,(3,((4,5),((6,8),7)))); 163
(1,(2,(3,((4,((6,7),8)),5)))); 145
(1,(2,((3,((6,8),7)),(4,5)))); 96
(1,(2,((3,(4,5)),((6,7),8)))); 66
(1,(2,(3,((4,(6,(7,8))),5)))); 51
```

```
(1,(2,((3,(4,5)),(6,(7,8))))); 22
(1,(2,(3,((4,((6,8),7)),5)))); 15
(1,(2,((3,(4,5)),((6,8),7)))); 1
```

Syntax and Options. To run `mbsum` for a single data file, type:

```
mbsum [-h] [--help] [-n #] [-o filename] [--version] <inputfilename(s)>
```

For example, let's say an alignment `mygene.nex` was analyzed with MrBayes with two runs, and sampled trees are in files `mygene.run1.t` and `mygene.run2.t`. These two sample files include, say, 5000 burnin trees each. To summarize these 2 runs use

```
mbsum -n 5000 -o mygene mygene.run1.t mygene.run2.t
```

or more generally

```
mbsum -n 5000 -o mygene mygene.run?.t
```

Here is a description of the available options.

`-h` or `[--help]`	prints a help message describing the options then quits.
`-n #` or `[--skip #]`	This option allows the user to skips lines of trees before actually starting the tally tree topologies. The default is 0, i.e *no* tree is skipped. The same number of trees will be skipped in each input file.
`-o filename` or `--out filename`	sets the output file name. A single output file will be created even if there are multiple input files. The tally combines all trees (except skipped trees) found in all files.
`--version`	prints the program's name and version then quits.

Example: the raw data and output from MrBayes are provided for the very first gene in the set analyzed in Ané *et al.* (2007). They are located in `$BUCKY_HOME/bucky/data/yeast/y000/`. The tree files from MrBayes, named `y000.run1.t` through `y000.run4.t`, each contain 5501 trees. They can be summarized with:

```
mbsum -n 501 -o y000.in    $BUCKY_HOME/bucky/data/yeast/y000/y000.run?.t
```

Warnings. `mbsum` will overwrite a file named `filename` if such a file exists. Only the first file is used to obtain the translate information, in case several runs are combined for a given gene. `mbsum` assumes an identical translate table in all separate runs. Translate tables are allowed by version 1.3.0 and required by version 1.3.2 and up.

From version 1.4.0, `bucky` and `mbsum` no longer assume that the same taxon is used to root all the trees across all runs and loci. Note also that MrBayes and BUCKy infer unrooted trees. Rooting is only used when writing trees in parenthetical format.

Running bucky

After input files created by `mbsum` are ready, the names of these files can either be given as arguments to `bucky`, or the file names can be written into a file, which in turn can be given to `bucky`. To run `bucky`, use either way:

```
bucky [-options] <summary_files>
bucky [-options]
```

With the second command, one of the options must be `-i filename`, where `filename` is the name of a file containing the list of all the input files (one input file per gene). For example, after creating all `.in` files with `mbsum` in the same directory, you can run bucky with the default parameters by typing this:

```
bucky *.in
```

Options.

`-i inputfilelist-file`	To give the list of files created by `mbsum` from a file.
`-o output-file-root`	This option sets the names of output files. Default is `run1`.
`-a alpha`	α is the *a priori* level of discordance among loci. Default α is 1.
`-n number-generations`	Use this option to increase the number of updates (default: 100,000). An extra number of updates will actually be performed for burnin. This number will be 10% of the desired number `n` of post-burning updates. The default, then, is to perform 10,000 updates for burnin, which will be discarded, and then 100,000 more updates.
`-h` or `--help`	Prints a help message describing options, then quits.
`-k number-runs`	Runs k independent analyses. Default is 2.
`-c number-chains`	Use this option to run Metropolis coupled MCMC (or MCMCMC), whereby hot chains are run in addition to the standard (cold) chain. These chains occasionally swap states, so as to improve their mixing. The option sets the total number of chains, including the cold one. Default is 1, i.e. no heated chains.
`-r MCMCMC-rate`	When Metropolis coupled MCMC is used, this option controls the rate r with which chains try to swap states: a swap is proposed once every r updates. Default is 100.
`-m alpha-multiplier`	Warm and hot chains, in MCMCMC, use higher values of α than does the cold chain. The cold chain uses the α value given by the option `-a`. Warmer chains will use parameters $m\alpha, m^2\alpha, \ldots, m^{c-1}\alpha$. Default m is 10. The independence prior corresponds to $\alpha = \infty$ so MCMCMC is not used with this prior.

`-s subsample-rate`	Use this option for thinning the sample. All post-burnin samples will be used for summarizing the posterior distribution of gene-to-tree maps, but you may choose to save just a subsample of these gene-to-tree maps. One sample will be saved every s updates. This option will have an effect only if option `--create-sample-file` is chosen. Default is 1: no thinning.
`-s1 seed1`	Default is 1234.
`-s2 seed2`	Default is 5678.
`-cf cutoff`	All splits with estimated sample-wide CF above this cutoff will be included in the list to have their summary information and their genome-wide CF displayed. Default is 0.05.
`-p file-with-prune table`	Default is to only consider taxa common to all genes, and prune all other taxa from the trees. This option allows the user to indicate which taxa should be retained in the analysis. These taxa are specified in a standard translate table, in a separate input file.
`-sg`	Use this option to skip processing genes that do not have all taxa of interest. Use `-p` to specify the taxa of interest.
`--calculate-pairs`	Use this option to calculate the posterior probability that pairs of loci share the same tree. Default is to NOT use this option.
`--create-sample-file`	Use this option for saving samples of gene-to-tree maps. Default is to NOT use this option: samples are not saved. Saving all samples can slow down the program.
`--create-joint-file`	This option creates a `.joint` file. NOT created by default.
`--create-single-file`	This option creates a `.single` file. NOT created by default.
`--use-independence-prior`	Use this option to assume *a priori* that loci choose their trees independently of each other. This is equivalent to setting $\alpha = \infty$. Default is to NOT use this option.
`--use-update-groups`	Use this option to permit all loci in a group to be updated to another tree. Default is to use this option, because it improves mixing.
`--do-not-use-update-groups`	Use this option to disable the update that changes the tree of all loci in a group in a single update. Default is to NOT use this option. If both options `--use-update-groups` and `--do-not-use-update-groups` are used, only the last one is applied. No warning is given, but the file `.out` indicates whether group updates were enabled or disabled.
`--opt-space`	This option accommodates large data sets that require large memory (many genes each with many unique trees), with space optimization. This option allows very large data sets to be analyzed, but increases the program running time.
`--genomewide-grid-size`	Number of grid points to estimate the genome-wide concordance factors. Default to 1000. Is automatically increased (with a message) to make sure it is greater than the number of sampled genes.

Output. Running `bucky` will create various output files. With defaults parameters, these output files will have names of the form `run1.*`, but you can choose you own root file name. The following output files describe the input data, input parameters, and progress history.

`.out`	Gives the date, version, input file names, parameters used, running time and progress history. If MCMCMC is used, this file will also indicate the acceptance history of swaps between chains.
`.input`	Gives the list of input files. There should be one file per locus.
`.single`	Gives a table with tree topologies in rows and loci in columns. The entries in the table are posterior probabilities of trees from the separate locus analyses. It is a one-file summary of the first step of BCA.

The following files give the full results as well as various result summaries. The goal of BCA this is to estimate the primary concordance tree. This tree is formed by all clades with concordance factors (CF) greater than 50%, and possibly other clades. The CF of a clade is the proportion of loci that have the clade. Sample-wide refers to loci in the sample and genome-wide refers to loci in the entire genome.

`.concordance`	Main output: this file first gives the primary concordance tree topology in parenthetical format and again the same tree with the posterior means of sample-wide CFs as edge lengths. A estimated population tree is also provided, inferred from the set of quartets with highest CFs. The list of clades in the primary concordance tree follows, with information on their sample-wide and genome-wide CFs: posterior mean and 95% credibility intervals. Inference on genome-wide CFs assumes that loci were sampled at random from an infinite genome. Finally, the file gives the posterior distribution of sample-wide CFs of all clades, sorted by their mean CF. In this list, CFs are expressed in number of loci instead of proportions.
`.cluster`	Gives the posterior distribution of the number of clusters, as well as credibility intervals. A cluster is a group of loci sharing the same tree topology. Loci in different clusters have different tree topologies.
`.pairs`	Gives an l by l similarity matrix, l being the number of loci. Entries are the posterior probability that two given loci share the same tree.
`.gene`	For each locus, gives the list of all topologies supported by the locus (index and parenthetical description). For each topology is indicated the posterior probability that the locus has this tree given the locus's data ('single' column) and given all loci's data ('joint' column).

`.sample`	Gives the list of gene-to-tree maps sampled by `bucky`. With n post-burnin updates and subsampling every s steps, this file contains n/s lines, one for each saved sample. Each line contains the number of accepted updates (to be compared to the number of genes * sub-sampling rate), the number of clusters in the gene-to-tree map (loci mapped to the same tree topology are in the same cluster), the log-posterior probability of the gene-to-tree map up to an additive constant followed by the gene-to-tree map. If there are l loci, this map is just a list of l trees. Trees are given by their indices. The correspondence between tree index and tree parenthetical description can be found in the `.gene` or `.single` file.
`.joint`	Gives a table with topologies in rows and loci in columns, similar to file `.single` file. Topologies are named by their indices as well as by their parenthetical descriptions. Entries are posterior probabilities (averaged across all runs) that each locus was mapped to each topology.

Examples

The example data provided with the program is organized as follows: directory `$BUCKY_HOME/bucky/data/example1/` contains 10 folders named `ex0` to `ex9`, one for each locus. These 10 folders contain a single file each, named `ex.in`, which was created by `mbsum`. For analyzing these data, one can use the default parameters and type

```
bucky $BUCKY_HOME/bucky/data/example1/ex?/ex.in
```

The question mark will match any character (any digit 0 to 9 in particular), so that all 10 locus files will be used for the analysis. The following command will run `bucky` with $\alpha = 5$, no MCMCMC, group updates disabled, 2 independent runs (default), one million updates and user-defined seeds (keep this command on a single line).

```
bucky -n 100000 -a 5 -s1 7452 -s2 9054 --do-not-use-update-groups
                                        $BUCKY_HOME/bucky/data/example1/ex?/ex.in
```

A look at the file `run1.concordance` shows that the clades (19,20) and (18,19) both have an estimated CF of 0.50 but that this estimate differed greatly between runs because its standard deviation is 0.707. Scrolling down the file indicates that the first run gave a 100% concordance factor to clade (19,20) all the time while the second run gave it a 0% concordance factor all the time. So the two runs are in very strong disagreement. These results could vary with different seeds. This poor mixing is fixed by using the option `--use-update-groups` (or by not using the `--do-not-use-update-groups` option!).

The yeast data analyzed in Ané *et al.* (2007) is provided with the program and organized as follows. The directory `$BUCKY_HOME/bucky/data/yeast/` contains 106 folders named `y000` to `y105`, one for each gene. In each of these folders, a file created by `mbsum` and named `run2.nex.in` contains the data from one gene. The list of all these input files is also provided, in `$BUCKY_HOME/bucky/data/yeast/yeast_inputfilelist` . For analyzing these data with

$\alpha = 2.5$, $n = 150,000$ updates, $k = 4$ independent runs, $c = 4$ chains (one cold and 3 hot chains), saving samples once every 1000 updates, and for keeping a similarity matrix among genes, one would type (on a single line)

```
bucky -a 2.5 -n 150000 -k 4 -c 4 --create-sample-file --calculate-pairs
               -s 1000   $BUCKY_HOME/bucky/data/yeast/y???/run2.nex.in
```

or alternatively, if `$BUCKY_HOME/bucky/` is the current directory:

```
bucky -a 2.5 -n 150000 -k 4 -c 4 --create-sample-file --calculate-pairs
               -s 1000 -i data/yeast/yeast_inputfilelist
```

To prune the analysis to a specific set of taxa, the option `-p` can be used with a file containing the taxon list like this:

```
bucky -n 150000 -p data/yeast/shortTaxonList -i data/yeast/yeast_inputfilelist
```

The above command aborts with an error message if some genes do not have all taxa of interest. To skip such genes and analyze the remaining genes only, `-sg` can be used. Example:

```
bucky -n 150000 -p data/yeast/shortTaxonList -sg -i data/yeast/yeast_inputfilelist
```

General notes

First step of BCA: Analysis of individual loci in MrBayes. Any model of sequence evolution can be selected for any locus: there need not be one model common to all loci. Some loci can be protein alignments, others DNA alignments, some can combine DNA and coded gap characters. Morphological characters could technically be included as one locus, but then the resulting concordance factors may not have an easy interpretation.

If hundreds of genes are to be analyzed, the analysis of these genes needs to be automated, and ideally run in parallel. One way to proceed is to have all the alignments in a single nexus file. In the first step, MrBayes can be told to ignore all but a single locus, and this would be repeated for each locus. Alternatively, MrBayes can be told to analyse all loci at once by unlinking *topologies, branch lengths* and other parameters assumed to be independent. The downside of this single analysis is the difficulty to use multiple processors in parallel to speed up the analysis.

Choosing the *a priori* level of discordance α. To select a value based on biological relevance, the number of taxa and number of genes need to be considered. For example, the user might have an *a priori* for the proportion of loci sharing the same genealogy. One can turn this information into a value of α since the probability that two randomly chosen loci share the same tree is about $1/(1+\alpha)$ if α is small compared to the total number of possible tree topologies. Also, the value of α sets the prior distribution on the number of distinct locus genealogies in the sample. Go to `www.stat.wisc.edu/~ane/bucky/prior.html` for instructions to visualize this distribution, as a tool for the choice of α.

Concordance tree and Population tree. The estimated primary concordance tree is currently provided as a fully resolved tree, possibly including clades that are in less than 50% of gene trees. The user might want to unresolve some of these clades, in case other conflicting clades have lower but similar concordance factors. Information on the credibility intervals of CFs can be used to decide if a clade in the concordance tree has a significantly higher CF than that of a conflicting clade.

The estimated population tree is inferred from quartet concordance factors. For each set of 4 taxa, there are 3 possible quartets. The quartet with the greatest CF is retained, and a tree is built from this set of quartets using the quartet-joining algorithm Xin *et al.* (2007). If concordance factors are estimated without error, then this quartet-based method is guaranteed to recover the true species tree if all discordance is due to the coalescent process along this tree. Degnan *et al.* (2009). On each branch in this tree, the CFs of all quartets in agreement with the 4-way partition defined by the branch are averaged together ($\bar{\text{CF}}$) An estimated branch length u is calculated in coalescent units, using the formula

$$u = -\log((1 - \bar{\text{CF}}) * 3/2)\,.$$

On branches with perfect agreement ($\bar{\text{CF}} = 1$) such as external branches, u is set to a maximum value of 10. This tree should be used with caution, for several reasons. It has yet to be tested in simulation and empirical studies. Uncertainty in estimated coalescent units of estimated average quartet CFs is not available, and this tree has yet to be refined in cases with more than one individual per population.

Missing sequences. If some loci have missing sequences, i.e. missing taxa, then rows of missing data (????) need to be inserted in place of the missing taxon's sequence. A more efficient way to deal with missing sequences will be implemented in a future version of `bucky`.

Visualization of output. Online tools are provided here to aid in the visualization of BUCKy output: `http://ane-www.cs.wisc.edu/buckytools/buckytools.php`. This includes visualization of the primary concordance tree, and of alternate trees that display alternate splits, chosen by the user. The user can also compare the credibitity intervals of conflicting splits to see if they are overlapping or not. The estimated population tree can also be visualized, among other things.

Version history

version 1.4.4 Fixes a bug with the estimation of genome-wide CFs, which caused a segmentation fault after the MCMC had finished, during the generation of the .concordance file. This bug occurred when the number of sampled genes was large (> 1000) and when the `-cf` option was used to get genome-wide information on splits with very low CFs.

version 1.4.3 Compiles under Mavericks (OS X 10.9). Option `-no-population-tree` turns off the population tree building, to save time and memory with large numbers of taxa.

version 1.4.2 `mbsum` can read in a wider range of tree file formats, including that produced by MrBayes 3.2.1.

version 1.4.0 A population tree is inferred, based on quartet concordance factors. The data reading time and the MCMC running time have been optimized. Option `-sg` was added, to easily ignore genes that do not have all taxa specified in the input taxon file. Option `-opt-space` was added to allow the analysis of very large data sets.

All trees are re-formatted internally using the last taxon as a root, to waive the requirement that all input files use the same root in their tree format. Note that this requirement is usually met if tree files are generated by MrBayes *and* if the same taxon appears first for all genes, because MrBayes uses the first taxon by default to root the trees. `bucky` and `mbsum` can parse rooted trees, but treat them as unrooted.

version 1.3.2 Option `-p` was added, to easily run analyses on subsets of taxa. Input files with different taxon sets are allowed, with the analysis reduced to their common taxa. Translate tables are required in all input files.

An important bug was fixed in the group update algorithm, to get correct sampling probabilities.

Concordance factors appear as proportions, not as numbers of genes, in the parenthetical description of the concordance tree.

version 1.3.1 Option `-cf` was added. `mbsum` recognizes Mac-style line breaks as well as Unix-style line breaks. A bug was fixed in `mbsum`, so that taxon numbers no longer need to range from 1 through Ntax.

The maximum number of tips is no longer limited to 32. The bug that occurred with exactly 32 taxa in assembling clades with high concordance factors into a concordance tree was fixed.

version 1.2b Independent runs are implemented, with information on the standard deviation of clade's CF across runs. A bug was fixed with the group update. The `-i` option was added, which can be particularly useful when thousands of genes are to be analyzed. Translate tables are used, so that taxa may appear in a different order for different genes (but the same taxon has to be serve as an outgroup consistently across genes). `bucky` uses the translate information if provided by the files created by `mbsum`, and makes some necessary checks and warnings.

version 1.2 The main output file (`.concordance`) contains the primary concordance tree in parenthetical format. It also displays a more detailed summary for all splits with mean concordance factor above 0.10. A bug was fixed in the list of splits belonging to the primary concordance tree. Inference on genome-wide concordance factors is included. The help message is improved with a better display of available options and default parameter values.

The following output files, deemed unnecessary, are no longer produced: `.gene`, `.top`, `.topologies` and `.splits`. Output file named `.genepost` in version 1.1 is now named `.gene` in version 1.2. Output files `.joint` and `.single` are not produced unless requested by the user.

References

Ané, C., B. Larget, D. A. Baum, S. D. Smith, and A. Rokas. 2007. Bayesian estimation of concordance among gene trees. Molecular Biology and Evolution **24**:412–426.

Ané, C., B. Larget, D. A. Baum, S. D. Smith, and A. Rokas. 2007. Erratum for: Bayesian estimation of concordance among gene trees. Molecular Biology and Evolution **24**:1575.

Baum, D. A. 2007. Concordance trees, concordance factors, and the exploration of reticulate genealogy. Taxon **56**:417–426.

Degnan, J. H., M. DeGiorgio, D. Bryant, and N. A. Rosenberg. 2009. Properties of consensus methods for inferring species trees from gene trees. Systematic Biology **58**:35–54.

Larget, B. R., S. K. Kotha, C. N. Dewey, and C. Ané. 2010. BUCKy: Gene tree/species tree reconciliation with Bayesian concordance analysis. Bioinformatics **26**:2910–2911.

Xin, L., B. Ma, and K. Zhang. 2007. A new quartet approach for reconstructing phylogenetic trees: Quartet joining method. In *Computing and Combinatorics*, number 4598 in Lecture Notes in Computer Science, pp. 40–50. Springer Berlin / Heidelberg.