Understanding the content of HyPhy's JSON output files

Stephanie J. Spielman

July 2018

Most standard analyses in HyPhy output results in JSON format, essentially a nested dictionary. This page describes the contents of each method's JSON output.

We note that these files are easily parsed for downstream use with standard scripting languages, e.g. using the `json` package in Python or `jsonlite` package in R.

Shared fields

All standard selection analyses will have the following top-level fields. Note that the key **`display order`** appears in may JSON fields. This key is used strictly for displaying results in HyPhy Vision and have no scientific meaning.

`analysis`

This field contains information about the analysis method of interest and is comprised of the following keys:

- **`info`**, Gives method name a brief statement of its intended use
- **`version`**, Method version
- **`citation`**, Method reference
- **`authors`**, names of primary authors of method
- **`contact`**, primary author to contact
- **`requirements`**, Data required in order to execute method

`input`

This field contains information about the inputted dataset being analyzed and is comprised of the following keys:

- **`file name`**, Full path to the name of user-inputted alignment file
- **`number of sequences`**, The number of sequences in the user-inputted alignment
- **`number of sites`**, The number of sites in the user-inputted alignment
- **`partition count`**, The number of user-specified data partitions to be used in the given analysis
- **`trees`**, The inputted trees (with user-supplied branch lengths, if applicable). This field is itself a dictionary containing all provided trees in newick format. The key for each tree corresponds to the data partition to which it belongs, starting from 0. For example, if there is one specified partition, this field might look like the following:

```
trees":{
    "0":"((taxon1: 0.5, taxon2: 0.5):0.5),taxon3:0.5)"
    }
```

Alternatively, for two specified partitions, this field might look like the following:

```
"trees":{
```

```
"0":"((taxon1: 0.5, taxon2: 0.5):0.5),taxon3:0.5)",
"1":"((taxon1: 0.1, taxon2: 0.1):0.1),taxon3:0.1)"
   }
```

fits

This field will contain information about each fitted model in the given analysis, and as such each method will return different fits. Most fit fields will contain the following fields:

- **Log Likelihood**, the log likelihood of the fitted model
- **estimated parameters**, the number of parameters estimated during likelihood optimization (i.e., not including empirically-determined parameters)
- **AIC-c**, the small-sample AIC for the fitted model
- **Rate Distributions**, the inferred rate distribution under the fitted model. Depending on the model, this field can refer to different rates. See each fit description for an explanation of the rates provided.

Within **fits**, all methods will contain a field **Nucleotide GTR**, and most will contain a field for **Global MG94xREV** (note that this is termed **MG94xREV with separate rates for branch sets** in methods RELAX and BUSTED, and termed **Baseline MG94xREV** in aBSREL). Each of these fields will contain the following additional fields:

- **Nucleotide GTR**
  - **Equilibrium frequencies**, a vector of nucleotide frequencies obtained empirically, in alphabetical order (A, C, G, T)
  - **Rate Distributions**, a dictionary of inferred nucleotide substitution rates under the GTR model. Note that, for all inferences, the rate **A->G** (and therefore also **G->A**) is constrained to equal 1.
- **Global MG94xREV** (or analogous field in RELAX, BUSTED, aBSREL)
  - **Equilibrium frequencies**, a vector of codon frequencies obtained using the CF3x4 estimator, in alphabetical order (AAA, AAC, AAG,..., TTT)
  - **Rate Distributions**, inferred ω rates under the fitted model. Content in this field is method-specific.

data partitions

This field provides information about the specified partitions for a given analysis. In this context, partition refers to the case where different sites evolve according to different trees. It does not refer to branch sets.

Keys enumerate partitions (starting from 0). Each partition contains the following fields:

- **name**, an automatic name determined by HyPhy for the given partition.
- **coverage**, a list of the sites to which the given partition corresponds.

branch attributes

This field provides information branch-level inferences. It contains a field for each partition (starting from 0) as well as an **attributes** field. Each partition's field further contains a dictionary for each node (taxa and internal nodes) in the data containing information about the branch. For each key seen per node, the **attributes** field defines its meaning, either **branch length** or **node label** (indicating meta information).

For example, consider this (fake, for explanation purposes) example **branch attributes** field:

```
"branch attributes":{
       "0":{
         "Node0":{
           "Nucleotide GTR":0,
           "Global MG94xREV":0
          },
         "Node1":{
           "Nucleotide GTR":0.1866611825064384,
           "Global MG94xREV":0.1982719556638508
```

```
        },
       "taxon1":{
         "original name": "taxon1",
          "Nucleotide GTR":0.0371579129290541,
         "Global MG94xREV":0.03929183706601039
        },
       "taxon2":{
         "original name": "taxon2",
         "Nucleotide GTR":0.01588385857974938,
         "Global MG94xREV":0.01711781392713551
        }
        "taxon3":{
      "original name":"taxon!!!3~~",
      "Nucleotide GTR":0.3679870611263812,
      "Global MG94xREV":0.4073912756908664
    }
      },
     "attributes":{
       "original name":{
         "attribute type":"node label",
         "display order":-1
        },
       "Nucleotide GTR":{
         "attribute type":"branch length",
         "display order":0
        },
       "Global MG94xREV":{
         "attribute type":"branch length",
         "display order":1
        }
      }
  }
```


Here, we see a single partition (0) with two internal nodes (Node0, Node1) and three tips (taxon1, taxon2, taxon3). All of these nodes contain the keys **`Nucleotide GTR`** and **`Global MG94xREV`** associated with numerical values. Looking into the **`attributes`** dictionary, we see that these keys correspond to the **`attribute type`** "branch length". Therefore, the values present in each node's dictionary represent the inferred branch length at that node under the given model. We also see the key **`original name`** for only nodes which are *tips*. This attribute, recorded as a "node label" in **`attributes`**, provides the original taxon name provided to HyPhy. In the event that there are forbidden characters in the provided name (i.e. for taxon3), HyPhy maps this forbidden name to an acceptable name. The original provided name will be recorded as an attribute in the node's dictionary.

`tested`

This field indicates whether each node (taxon and internal node) belongs to either the "test" or "background" branch sets. If multiple partitions were specified, then there will be a dictionary for each partition, beginning from 0. In the case of RELAX, this field will indicate if branches are Test, Reference, or Unclassified.

`timers`

This field provides the run times, including total execution time (**`Total time`**), of different stages in model fitting. Each method will report the total time as well as times for critical fitting stages specific to the method.

BUSTED

This section details JSON fields which are specific to BUSTED, and further clarifies the contents of shared fields as they appear in BUSTED.

`background`

This field simply contains the value **0** or **1** indicating if all nodes are considered as test ("foreground"), or if specific test/background sets have been specified. **0** indicates that all branches are test, and **1** indicates that there are separate test and background lineages.

`test results`

This field reports the likelihood ratio test statistic (**`LRT`**) and P-value (**`p-value`**) obtained under the BUSTED test for gene-wide episodic selection.

`fits`

In BUSTED, this field contains either 3 or 4 model fits:

- **`Nucleotide GTR`**, whose contents are described above in the section **Shared fields**
- **`MG94xREV with separate rates for branch sets`**, the MG94xREV fit in BUSTED.
  - **`Rate Distributions`** reports the branch-set–wide inferred ω ratio. This field will contain either one or two keys (depending on if test and background were specified, or just test). For example, this field would contain the following contents for an analysis with test and background specified, where each list represente[s] [**dN, dS**] (as BUSTED does not consider synonymous rate variation, dS=1 in all cases):

```
"Rate Distributions":{
    "non-synonymous/synonymous rate ratio for *background*":[
            [0.5210400433507286, 1]
        ],
    "non-synonymous/synonymous rate ratio for *test*":[
            [0.434676603885773, 1]
        ]
    }
```

- **`Unconstrained model`**, the fitted BUSTED alternative model that allows for positive selection in test branches.
  - **`Rate Distributions`** reports the three-rate ω distribution inferred for the test lineages, as well as the background lineages if they were specified. Values for each rate category (numbered 0, 1, 2) include the inferred rate (**`omega`**) and the inferred proportion of sites evolving at this rate (**`proportion`**).
- **`Constrained model`**, the fitted BUSTED null model that disallows positive selection in test branches. Note that this field will only appear in the JSON **if there was evidence suggetive of selection in the earlier unconstrained model fit**. If, for example, the unconstrained model did not detect a proportion of sites with $\omega > 1$, BUSTED will skip the null fit and conclude that there is no evidence for selection in the test branches.
  - **`Rate Distributions`** reports the three-rate ω distribution inferred for the test lineages, as well as the background lineages if they were specified. It follows the same format as does the corresponding field in **`Unconstrained model`**.

`branch attributes`

The branch attributes in BUSTED consist of the following:

- **`original name`**, provided taxon names
- **`Nucleotide GTR`**, branch lengths under this model
- **`MG94xREV with separate rates for branch sets`**, branch lengths under this model
- **`unconstrained`**, branch lengths under the BUSTED alternative model

- **constrained**, branch lengths under the BUSTED null model, if it was fit during the given BUSTED analysis

Site Log Likelihood

Log-likelihood values calculated for each site, under the **unconstrained** and, if fit, the **constrained** models. A description of these values is available in the BUSTED publication and on hyphy.org.

Evidence Ratios

Site-wise evidence ratios providing descriptive evidence for whether each site might be selected. A description of these values is available in the BUSTED publication and on hyphy.org. Note that this field will **only be populated** if the Constrained model was fit.

aBSREL

This section details JSON fields which are specific to aBSREL, and further clarifies the contents of shared fields as they appear in aBSREL.

test results

This field contains three keys:

- **P-value threshold**, the specified p-value threshold for calling lineages as selected
- **tested**, the number of lineages tested for selection
- **positive test results**, the number of lineages inferred to have experienced positive selection

Lineage-specific test results are contained the in **branch attributes** field, described below.

fits

In aBSREL, this field contains 3 model fits:

- **Nucleotide GTR**, whose contents are described above in the section **Shared fields**
- **Baseline MG94xREV**, the MG94xREV fit in BUSTED.
  - **Rate Distributions** reports the distribution of branch-specific ω values inferred under this model, including the mean, median, and 95% bounds. The precise lineage ω estimates are contained in the **branch attributes** field, described below. An example Rate Distribution will look like this:

```
"Rate Distributions":{
   "Per-branch omega":{
     "Mean":714285715.2178631,
     "Median":0.6868187444722127,
     "2.5%":0,
     "97.5%":3.921475959531533
    }
 }
```

- **Full adaptive model**, the aBSREL adaptive model fit. Note that it will contain an *empty* **Rate Distributions** field. The inferred lineage ω distributions are contained in the **branch attributes** field, described below.

`branch attributes`

The branch attributes in aBSREL consist of the following:

- **`original name`**, provided taxon names
- **`Nucleotide GTR`**, branch lengths under this model
- **`Baseline MG94xREV`**, branch lengths under this model
- **`Baseline MG94xREV omega`**, branch-specific ω inferred under this model
- **`Rate classes`**, number of adaptively-inferred rate classes along this branch, indicating adaptive model complexity
- **`Full adaptive model`**, the branch lengths under this model
- **`Rate Distributions`**, the rate distribution (each rate in the form **[dN,dS]**) inferred under the adaptive model. There will be as many dN/dS ratios here as there are rate classes.
- **`LRT`**, the likelihood ratio test statistic for selection along this branch. This value will be 0 if there was insufficient evidence to warrant testing.
- **`Uncorrected P-value`**, the uncorrected p-value for selection along this branch. This value will be 1 if there was insufficient evidence to warrant testing.
- **`Corrected P-value`**, the Bonferroni-Holm corrected p-value for selection along this branch. This value will be 1 if there was insufficient evidence to warrant testing. **This p-value should be used as evidence for selection**.

RELAX

This section details JSON fields which are specific to RELAX, and further clarifies the contents of shared fields as they appear in RELAX.

`fits`

In RELAX, this field contains either either 4 or 6 model fits, where **`General descriptive`** and **`RELAX partitioned descriptive`** are only present when "All" RELAX models are fitted (in contrast to "Minimal" mode).

- **`Nucleotide GTR`**, whose contents are described above in the section **Shared fields**
- **`MG94xREV with separate rates for branch sets`**, the MG94xREV fit in RELAX.
  - **`Rate Distributions`** reports the branch-set–wide inferred ω ratio. This field will contain either two or three keys (always for Reference and Test, and additional Unclassified if this branch set exists). Each rate list shown here represents **[dN, dS]** (as RELAX does not consider synonymous rate variation, dS=1 in all cases).
- **`General descriptive`**, a three-rate MG94xREV fit for all branch sets combined. This model fits a separate **K** parameter for each branch. This information is available in the **`**branch attributes**`** field, described below.
  - **`Rate Distributions`** reports the the three-rate ω distribution shared across all branch sets. Values for each rate category (numbered 0, 1, 2) include the inferred rate (**`omega`**) and the inferred proportion of sites evolving at this rate (**`proportion`**).
- **`RELAX alternative`**, the RELAX alternative model fit.
  - **`Rate Distributions`** reports the the three-rate ω distribution for each branch set. Values for each rate category (numbered 0, 1, 2) include the inferred rate (**`omega`**) and the inferred proportion of sites evolving at this rate (**`proportion`**).
- **`RELAX null`**, the RELAX null model fit.
  - **`Rate Distributions`** reports the the three-rate ω distribution for each branch set. Values for each rate category (numbered 0, 1, 2) include the inferred rate (**`omega`**) and the inferred proportion of sites evolving at this rate (**`proportion`**).
- **`RELAX partitioned descriptive`**, a three-rate MG94xREV fit per branch set. Note that this model does not fit a **K** parameter.
  - **`Rate Distributions`** reports the three-rate ω distribution inferred for the all specified branch sets. Values for each rate category (numbered 0, 1, 2) include the inferred rate (**`omega`**) and the inferred proportion of sites evolving at this rate (**`proportion`**).

`test results`

This field reports the likelihood ratio test statistic (`LRT`) and P-value (`p-value`) obtained under the RELAX test. It additionally reports the inferred K parameter in the key `relaxation or intensification parameter`.

`branch attributes`

The branch attributes in RELAX consist of the following:

- `original name`, provided taxon names
- `Nucleotide GTR`, branch lengths under this model
- `MG94xREV with separate rates for branch sets`, branch lengths under this model
- `General descriptive`, branch lengths under this model. This attribute is only present if RELAX is run in "All" mode.
- `k (general descriptive)`, the branch-specific **K** parameter inferred under the General Descriptive model. This attribute is only present if RELAX is run in "All" mode.
- `RELAX alternative`, branch lengths under this model.
- `RELAX null`, branch lengths under this model.
- `RELAX partitioned descriptive`, branch lengths under this model. This attribute is only present if RELAX is run in "All" mode.

`tested`

This field indicates whether each node (taxon and internal node) belongs to either the "Test", "Reference", or (if specified) "Unclassified" branch sets.

FEL and MEME

FEL and MEME JSONs contain very similar contents, most of which is described under **Shared fields** above. This section details JSON fields in FEL and MEME.

`branch attributes`

The branch attributes in FEL and MEME consist of the following:

- `original name`, provided taxon names
- `Nucleotide GTR`, branch lengths under this model
- `Global MG94xREV`, branch lengths under this model

`MLE`

This field contains all site-level estimates of selection and effectively corresponds to a csv of inferences. It contains two keys:

- `headers`, the header/meaning of each column in the `content` block
- `content`, site-level estimates of selection, with values ordered as in `headers`. Each row corresponds to a site, in the order the sites were listed in `data partitions`. There will be a separate `content` section for each partition, starting from 0.

LEISR

The LEISR JSON is very similar to those of FEL and MEME. This section details JSON fields in LEISR.

fits

This field contains a description of the specified model fit. As such, the specific name will differ based on what model the user has selected.

- **<model>**
  - **Rate Distributions** is either empty if rate heterogeneity is off, or contains information about the rate distribution

Note that in version <=2.3.7, there was a minor bug in the LEISR JSON, such that the field **Equilibrium Frequencies** in **fits** was misnamed as **EFV**. This has been addressed for HyPhy version >=2.3.8.

branch attributes

The branch attributes in LEISR consist of the following:

- **original name**, provided taxon names
- **<model>**, branch lengths under the given model name in the **fits** field

MLE

This field contains is analogous to those of FEL and MEME. Note that only a single partition is allowed for LEISR inference at this time.

SLAC

The SLAC JSON is very similar to those of FEL and MEME but contains some additional information. This section details JSON fields in SLAC.

branch attributes

The branch attributes in SLAC consist of the following:

- **original name**, provided taxon names
- **Nucleotide GTR**, branch lengths under this model
- **Global MG94xREV**, branch lengths under this model
- **codon**, the codon sequence, presented as a list, at this node. For internal nodes, these codon sequences were inferred using maximum likelihood ancestral state reconstruction.
- **amino-acid**, the amino acid sequence, presented as a list, at this node. For internal nodes, these sequences were inferred using maximum likelihood ancestral state reconstruction.
- **synonymous substitution count**, the number of synonymous substitutions per site inferred to have occured along this branch
- **nonsynonymous substitution count**, the number of nonsynonymous substitutions per site inferred to have occured along this branch

MLE

This field is arranged similar to MEME and FEL, with **headers** representing a csv header and **content** representing corresponding rows under this header, per partition. However, it contains additional content information as well:

- **content** contains both **by-site** and **by-branch** inferences (that is, substitution information across sites and across branches) under both **RESOLVED** and **AVERAGED** ancestral state modes. The **by-branch** inferences will contain the field **NAMES** indicating the order that rows within are presented.

- **sample-median** contains **RESOLVED** and **AVERAGED** versions of the header contents referring to the median of the bootstrap draws for establishing a 95% confidence interval.
- **sample-2.5** contains **RESOLVED** and **AVERAGED** versions of the header contents referring to the lower 2.5% boundary of the bootstrap draws for establishing a 95% confidence interval.
- **sample-97.5** contains **RESOLVED** and **AVERAGED** versions of the header contents referring to the upper 97.5% boundary of the bootstrap draws for establishing a 95% confidence interval.

FUBAR

This section details JSON fields which are specific to FUBAR, and further clarifies the contents of shared fields as they appear in FUBAR.

MLE

This field follows the same format as in FEL/MEME. Please see the **FEL and MEME** section above for details.

settings

This field records the run-time settings specified for FUBAR inference:

- **grid size**, size of the NxN grid used to pre-compute rates and likelihoods
- **chains**, the number of MCMC chains run
- **chain-length**, the number of generations run per chain
- **burn-in**, the number of samples discarded per chain as burn-in
- **samples**, the number of samples drawn per chain
- **concentration**, the alpha parameter for thr Dirichlet prior
- **posterior**, the posterior probalility threshold used to call selected sites

grid

This field contains the full grid of rates and their corresponding weights, where each row represents the following:

```
[nonsynonymous rate (dN), synonymous rate (dS), posterior mean over the grid for this dN/dS rate]
```

posterior

This field contains, for each partition (starting from 0), the posterior probability for each grid rate across all sites (starting from 0).

branch attributes

The branch attributes in FUBAR consist of the following:

- **original name**, provided taxon names
- **Nucleotide GTR**, branch lengths under this model

FADE

This section details JSON fields which are specific to FADE, and further clarifies the contents of shared fields as they appear in FADE.

MLE

This field is arranged similarly to MEME/FEL, with headers representing a csv header and content representing corresponding rows under this header. In FADE, the **content** field is organized per amino acid, with each amino-acid field containing a separate array for each partition. For example, the **A** field nested within **content** provides parameters estimated during FADE's test for directional evolution towards alanine at each site in each partition. Importantly, the final column in the content gives the Bayes Factor indicating whether the site has evolved under directional selection towards the amino acid of interest, where values over 100 provide strong statistical evidence.

settings

This field records the run-time settings specified for FADE inference:

- **grid size**, size of the NxN grid used to pre-compute rates and likelihoods
- **chains**, the number of MCMC chains run
- **chain-length**, the number of generations run per chain
- **burn-in**, the number of samples discarded per chain as burn-in
- **samples**, the number of samples drawn per chain
- **concentration**, the alpha parameter for thr Dirichlet prior
- **bayes factor**, the Bayes Factor threshold used to call directionally selected sites
- **method**, the algorithm used during posterior inference

site annotations

This field is arranged similarly to MEME/FEL, with headers representing a csv header and content representing corresponding rows under this header, per partition. In FADE, this field contains information regarding the subsitution history of each site, with headers representing a csv header and content representing corresponding rows under this header, per partition. The two columns are as follows:

- **Composition** provides the amino-acid composition, *including inferred ancestral states*, of the given site along all branches (not just the selected foreground). This column will contain an empty string `""` if the site contains only missing and/or amibguous states. For example, an entry of `"A37,P1"` indicates that the site column *and* inferred internal node states contain, in total, 37 alanines and 1 proline.
- **Substitutions** provides the inferred amino-acid substitutions of the given site along *foregound* branches only. This column will contain an empty string `""` if the site has not experienced any substitutions, according to ancestral reconstruction under the specified model. Substitutions are represented by text arrows, with commas separating different starting amino-acids. For example, an entry of `"T->A(2)I(2)L(1)V(1), V->G(1)"` indicates the following:
  - Two substitutions from threonine to alanine
  - Two substitutions from threonine to isoleucine
  - One substitution from threonine to leucine
  - One substitution from threonine to valine
  - One substitution from valine to glycine

branch attributes

The branch attributes in FUBAR consist of the following:

- **original name**, provided taxon names
- **<model>**, branch lengths under the given protein model. The name of this field will correspond to the model name provided to FADE analysis, which is also shown in **fits**.