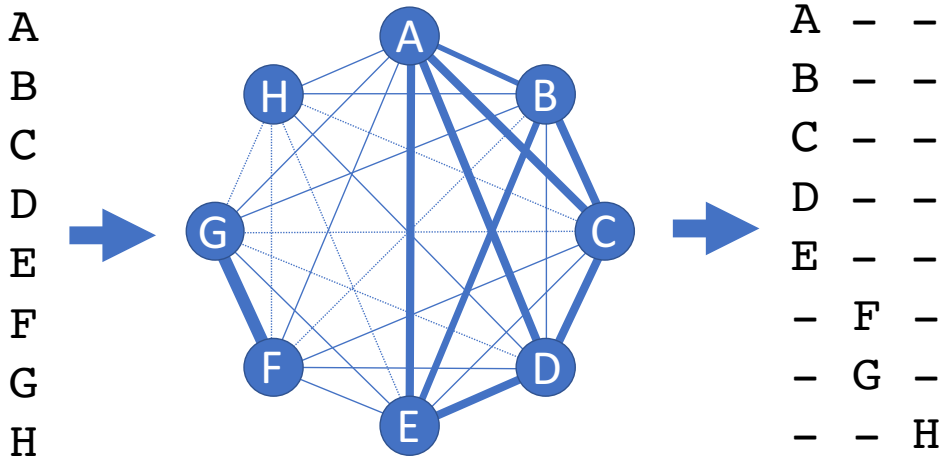
Divvier the Manual

Divvier is a program intended for the processing of multiple sequence alignments (MSA) before downstream analysis. It was primarily developed for phylogenetic inference, but the core concepts might be applicable to any application where statistical evidence of shared ancestry is important. The manual should provide an overview of the methodology – including its strengths and weaknesses – and how to apply it to your data.

Frequently Asked Questions

What does Divvier do?

Divvier uses a probabilistic model, based on the substitution models and indel models used in molecular evolution, to estimate the posterior probabilities of pairs of residues sharing an ancestor in a multiple sequence alignment (MSA). It then evaluates these posterior probabilities one column at a time to cluster together residues that show evidence of shared homology. To clarify this look at the figure below



The left hand side shows a column containing residues A-H. The center represents the graph with lines showing all the pairwise links between the homologies, with the thickness of the lines representing the posterior probabilities. In this case the evidence suggests there should be three groups {A,B,C,D,E}, {F,G} and {H}. The right hand side shows that these can be represented by a block of three new columns in the MSA, each representing a group of residues with strong evidence for a common ancestor. Divvier does this for every column in the original MSA and returns a *divvied* or *partially filtered* MSA.

How do I cite Divvier?

At the moment Divvier is submitted and can be cited as:

R.H. Ali, M. Bogusz and S. Whelan. Splitting Alignments. A graph-based approach for improving the homology inference in multiple sequence alignments. *Submitted*.

What's this about Divvying and Partially Filtering?

Divvying and partial filtering are two different ways of treating the block of columns shown on the right hand side of the figure above. Divvying returns the whole block and keeps all of the sequence information intact and just removes low confidence homologies whereas partial filtering only returns the largest cluster, in this case {A,B,C,D,E}, which retains the MSA length and removes some individual residues, leading to some columns being partially filtered.

What do you mean by evidence of shared homology?

We use shared (pairwise) homology to refer to the case where two or more residues have high posterior probabilities linking them together. For instance the thick line linking F and G in the figure above suggests the have shared pairwise homology. When considering groups of residues, such as {A,B,C,D,E}, we use a clustering method to decide which residues belong together. This clustering method does not require all residues to show very strong evidence of homology between all pairs within the cluster – for instance B and D have a low posterior probability – but instead assesses high average support within the cluster. Note that this is a statistical method so some true homologies might show low evidence of shared homology (false negative) and some false homologies might show high evidence of shared homology (false positive). These errors are inevitable in any difficult problem and our results suggest that Divvier does a better job of identifying true and false positives than other existing methods (see the paper for full details).

Should I believe your posterior probabilities

These posterior probabilities are derived from a probabilistic model and you should only trust them as much as you trust the model. Our results suggest this model does a reasonable job at identifying true and false positives, but we cannot say it has the best performance possible. We have, however, tried several other models and the posterior probabilities seem pretty robust to model choice. This might not be that surprising since most models try and capture the same information about the evolutionary process.

When do I need a good divvying/filtering method?

There seem to be two main factors deciding whether divvying or filtering might be important: sequence divergence and intended purpose of the MSA. In general the MSA problem becomes harder as the sequence divergence increases, so anything that can identify and account for this uncertainty at least has the potential to help downstream analyses. The degree that results are affected by small errors in the MSA might also affect your decision to divvy or filter your data. Some methods, such as inferring the selective pressures acting on proteins using PAML, seem rather sensitive even to small errors in the MSA, so might be improved by applying partial filtering. Note we have not directly examined this effect, so your mileage may vary.

Does divvying or filtering always improve results?

Our results suggest that our divvying or partial filtering methods do help downstream analyses, although we cannot guarantee this effect to hold for all future cases. Users should note there is some evidence that filtering does not help downstream analyses [10.1093/sysbio/syv033].

How do you recommend I run divvier?

There are two main modes for running divvier on an alignment file: full divvying and partial filtering. Full divvying is the default option and will aim to retain as much information in the MSA as possible. For immediate phylogenetic use we suggest the following that will require at least 4 character per column and produce output to myfile.divvy.fas that will work as input into standard phylogeny programs:

```
./divvier -mincol 4 -divvygap myfile.fas
```

Partial filtering can be done by adding a single option to this command line:

```
./divvier –partial -mincol 4 -divvygap myfile.fas
```

When should I run divvier?

You should run divvier directly after performing the MSA. It requires the full sequences to make reliable statements of confidence in the pairwise homologies, so it cannot be safely run after another filtering program.

I've just divvied an MSA and it looks horrible: HELP!

Divvying (and partial filtering) only allows columns that have strong evidence of shared homology. This approach can introduce a lot of gaps when there is a lot of MSA uncertainty, so that's what's happened. If you want nicer looking alignments tune the -mincol setting to ensure you're keeping a certain number of characters. If you're really still uncomfortable then try partial filtering instead, which is likely to produce an MSA closer to that from a classical filtering method.

Do you recommend partial filtering or divvying?

Our results suggest both partial filtering and divvying provide a substantial improvement over existing filtering methods and that divvying can provide a small improvement of partial filtering. For some, partial filtering might be a preferable option since it produces shorter MSAs that are easier and faster to analyze.

How do I use divvier is phylogenomic studies

Divvier measures the statistical support for pairwise homologies that are based on comparisons of pairs of sequences. This approach means that you need to run divvier on the separate gene-based MSAs before you concatenate them otherwise these measures of certainty will be screwy and likely give you bad results.

What does divvier not do?

Data derived errors in phylogeny can be broadly broken into three categories (at least from the alignment perspective): i) input errors, where parts of the sequence are not really homologous; ii) MSA uncertainty errors, where incorrect homologies can affect evolutionary parameters such as the tree or selection; and iii) modeling errors, where the evolutionary model cannot adequately capture the complexity of evolution, such as site-wise or lineage-wise heterogeneity in composition or rate. Divvier only explicitly addresses MSA uncertainty errors. Divvier might help with input errors, but we recommend you use the program PREQUAL to identify and remove these errors. Divvier will in no way help with modeling errors.

Divvier options

Clustering options

Divvier uses a UPGMA-type clustering method to identify clusters of putative true homologies. The following options affect that clustering:

`–divvy`

Default and will make divvier run full divvying.

`–partial`

Make divvier run in partial filtering mode.

`–threshold`

Adjust the threshold in the clustering step for deciding evidence of pairwise homology. High values are stricter (fewer false positives) and low values are more liberal (more false positives). The default values (divvying = 0.804; partial 0.774) were determined from a 1% false discovery rate from the BAliBASE database and are found to be effective for most cases since even a small number false positive homologies (MSA errors) cause unpredictable parameter estimates.

Output options

There are several options to make output easier to work with in divvier.

-mincol X

Specifies the minimum number of characters there must be in a column for it to be output. The default is 2 since that's the first time there will be phylogenetic information in the column, but an appropriate alternative might be 4 since that's the first time a column might help determine a tree. In general higher values will yield shorter MSAs, which might help computation. Works for both divvying and partial filtering.

-divvygap

By default divvier outputs a static '*' character when a column is divvied. This allows visual inspection of the divvying and allows the reconstitution of columns. If you chose this option it will just output a gap '-' character, so the divvied MSA can be put directly into a phylogenetic program, such as RAxML or IQ-TREE.

Approximation options

Divvier uses a lot of heuristics to perform fast calculations. These have been extensively tests on BAliBASE and you probably don't want to change these options, but in case you do:

`-approx X`

Divvier avoids calculating an all against all comparison of sequences since this can be slow for moderate to large data sets (>20 sequences). Instead it computes a subset of comparisons for each considered cluster. The -approx X option adjusts the X number of sequences in that subset. Larger numbers of X will tend to be more accurate and X >= number of sequences will be exhaustive.

`-checksplits`

Some columns might not have information for every homology, especially in sparse columns. When this happens you'll see "WARNING: some columns had no information supporting or refuting divvying clusters". Divvier is conservative in these cases and rejects those homologies, but this option overrides that rejection and ensures all comparisons are covered. It can be very slow for larger numbers of sequences.

`-HMMapprox` (default)

`-HMMexact`

The probability model (pairHMM) uses a bounded approach in the dynamic programming matrix to reduce computation time. This option is specified by –HMMapprox and can be turned off by using -HMM exact