

SUMAC: Supermatrix Constructor version 1.0

Manual

William A. Freyman¹

Department of Integrative Biology, University of California,
Berkeley

¹freyman@berkeley.edu

Contents

1	Introduction	3
1.1	About SUMAC	3
1.2	Installation	4
1.2.1	Requirements	4
1.2.2	Installing Requirements on Linux	4
1.2.3	Installing Requirements on Mac	5
1.2.4	Installing SUMAC	5
1.3	License and Warranty	5
2	Quick Start Tutorial	6
2.1	Construct a Supermatrix	6
2.2	Explanation of the Output Files	7
3	SUMAC in Detail	9
3.1	Downloading GenBank	9
3.1.1	GenBank Division	9
3.1.2	GenBank File Path	9
3.2	Specifying Ingroup and Outgroup	9
3.3	Using Guide Sequences	9
3.4	Partial Decisiveness	10

3.5	Supermatrix Figure	10
-----	------------------------------	----

Chapter 1

Introduction

1.1 About SUMAC

SUMAC (Supermatrix Constructor) is a Python package to facilitate the construction and evaluation of phylogenetic supermatrices. It is designed to be run as a command-line program, though the modules can also be imported and used in other Python scripts. SUMAC will assemble supermatrices for any taxonomic group recognized in GenBank, and is optimized to run on multicore processors and clusters by utilizing multiple parallel processes.

When run from the command-line, SUMAC will perform a number of steps to create the phylogenetic supermatrix. First, SUMAC will download the GenBank database for the specified GenBank division (PLN, MAM, etc). SUMAC will then build clusters of homologous sequences in one of two ways: (1) perform exhaustive all-by-all BLAST comparisons of each ingroup and outgroup sequence and use a single-linkage hierarchical clustering algorithm, or (2) BLAST each ingroup and outgroup sequence against user provided guide sequences that define each cluster. SUMAC then discards clusters that are not phylogenetically informative (< 4 taxa), and then aligns each cluster of sequences using MUSCLE. Finally, the align-

ments are concatenated by species name (using the GenBank taxonomy) creating a supermatrix. A number of metrics are then calculated on the supermatrix, a graph indicating taxon coverage density is generated, and spreadsheets (in CSV format) are produced with information about each DNA region and GenBank accession used in the supermatrix. There are many options described in detail in later chapters of this manual.

1.2 Installation

1.2.1 Requirements

The following requirements must be installed to run SUMAC:

Python 2.7

Biopython

MUSCLE

BLAST+

1.2.2 Installing Requirements on Linux

The following commands install the requirements for Debian GNU/Ubuntu Linux systems:

```
git clone https://github.com/biopython/biopython.git
cd biopython
python setup.py build
python setup.py test
sudo python setup.py install
sudo apt-get install ncbi-blast+
sudo apt-get install muscle
```

TODO: use pip to install Biopython??

1.2.3 Installing Requirements on Mac

TODO!

1.2.4 Installing SUMAC

TODO: use pip to install...

```
python setup.py install
```

1.3 License and Warranty

SUMAC is free software; you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation; either version 3 of the License, or (at your option) any later version.

The program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details (<http://www.gnu.org/copyleft/gpl.html>).

Chapter 2

Quick Start Tutorial

This chapter provides the quick way to get started using SUMAC. There are many details described in Chapter 3 that would likely be helpful.

2.1 Construct a Supermatrix

The most basic usage of SUMAC is to build a supermatrix with the following command:

```
python -m sumac -d pln -i Onagraceae -o Lythraceae
```

This command will download the PLN (Plant) GenBank division and search it for all sequences within the taxonomic groups Onagraceae (the ingroup) and Lythraceae (the outgroup). SUMAC will then perform all-by-all BLAST comparisons of each sequence, build clusters of putatively homologous sequences, and construct a supermatrix.

Unless you are on a large multi-core system, the all-by-all BLAST comparisons will take a very long time to be performed since well over 5000 sequences will be found. To speed up the supermatrix construction, you could make a FASTA file of

guide sequences to define each cluster. Each guide sequence could be an example of a sequence commonly used for phylogenetic analysis. You could then use this command:

```
python -m sumac -d pln -i Onagraceae -o Lythraceae -g guides.fasta
```

Which approach is better for constructing supermatrices? Using guide sequences makes supermatrix construction much faster, however it requires a priori knowledge of which DNA regions will be used in the supermatrix. Performing all-by-all BLAST comparisons is computationally more expensive, but it effectively data-mines GenBank in an exploratory fashion, so that sequence data not necessarily used in previous systematic studies can also be incorporated into the supermatrix. The decision will depend on the size of the taxonomic group being analyzed and the computational resources available.

2.2 Explanation of the Output Files

SUMAC will output the following files:

```
alignments/combined.fasta
alignments/N.fasta
clusters/N.fasta
gb_search_results
genbank_accessions.csv
gene_regions.csv
plot.pdf
sumac_log
```

The `alignments/combined.fasta` is the final aligned supermatrix in FASTA format. `alignments/N.fasta`, where N is an integer > 0 , is the alignment of gene

region N . Similarly, `clusters/N.fasta` is the unaligned raw sequence cluster of gene region N . The `gb_search_results` file is used by SUMAC to save the results of the GenBank sequence search in case the search is re-run. This file is not human readable.

The two `CSV` (comma-separated values) files contain tables that provide useful summary information about the supermatrix. `genbank_accessions.csv` is a table with each GenBank accession used, ordered by gene region and taxon (like the appendices found in most systematics papers). The `gene_regions.csv` file contains the number of taxa, the aligned length, the percent missing data, and the taxon coverage density of each gene region used in the supermatrix.

The `plot.pdf` file is a figure that shows how much sequence data was available for each taxon for each gene region. `sumac_log` is a log of the SUMAC run, and contains a great deal of information about the supermatrix construction, including final metrics such as the partial decisiveness (PD) of the supermatrix.

Chapter 3

SUMAC in Detail

3.1 Downloading GenBank

3.1.1 GenBank Division

blah blah

3.1.2 GenBank File Path

blah

3.2 Specifying Ingroup and Outgroup

blah b

3.3 Using Guide Sequences

blah blah

3.4 Partial Decisiveness

blah blah

3.5 Supermatrix Figure

blah blah