CHAPTER 1

Build information
CHAPTER 2

PyPI version and Licensing
Chapter 2. PyPI version and Licensing
If you’re feeling impatient, please head over to the *QuickStart Guide*
**pyani** is a program and Python package that provides support for calculating average nucleotide identity (ANI) and related measures for whole genome comparisons, and for rendering relevant graphical and tabular summary output. Where available, it natively takes advantage of multicore systems, and can integrate with SGE or OGE-compatible job schedulers to manage the computationally-heavy sequence comparisons.

Installing the **pyani** Python package also installs the program **pyani**, which enables command-line based analysis of genomes. Results are stored in a private SQLite3 database local to the machine, which permits addition of genomes to a previous analysis without having to recalculate all previously-performed comparisons, facilitating incremental analysis and visualisation, and enabling incremental maintenance of results for a taxonomic group as new genome sequences become available.
CHAPTER 5

Reporting problems and requesting improvements

If you encounter bugs or errors, or would like to suggest ways in which pyani can be improved, please raise a new issue at the pyani GitHub issues page.

If you’d like to fix a bug or make an improvement yourself, contributions are welcomed, and guidelines on how to do this can be found at the Contributing to pyani documentation page.

5.1 Citing pyani

We would be grateful if you could please cite the following manuscript in your work if you have found pyani useful:


@Article{C5AY02550H,
author ="Pritchard, Leighton and Glover, Rachel H. and Humphris, Sonia and Elphinstone, John G. and Toth, Ian K.",
title ="Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens",
journal ="Anal. Methods",
year ="2016",
volume ="8",
issue ="1",
pages ="12-24",
publisher ="The Royal Society of Chemistry",
doi ="10.1039/C5AY02550H",
url ="http://dx.doi.org/10.1039/C5AY02550H",
abstract ="Soft rot Enterobacteriaceae (SRE) are bacterial plant pathogens that cause blackleg, wilt and soft rot diseases on a broad range of important crop and ornamental plants worldwide. These organisms (spanning the genera Erwinia, Pectobacterium, Dickeya, and Pantoea) cause significant economic and yield losses in the field, and in storage. They are transmissible through surface water, by trade and other movement of plant material and soil, and in some cases are (continues on next page)"
subject to international legislative and quarantine restrictions. Effective detection —and diagnosis in support of food security legislation and epidemiology is dependent on the ability to classify —pathogenic isolates precisely. Diagnostics and classification are made more difficult by the influence of horizontal gene transfer —on phenotype[,] and historically complex and sometimes inaccurate nomenclatural and taxonomic assignments that persist in —strain collections and online sequence databases. Here[,] we briefly discuss the relationship between taxonomy[,] genotype —and phenotype in the SRE[,] and their implications for diagnostic testing and legislation. We present novel whole-genome —classifications of the SRE[,] illustrating inconsistencies between the established taxonomies and evidence from —completely sequenced isolates. We conclude with a perspective on the future impact of widespread whole-genome sequencing and —classification methods on detection and identification of bacterial plant pathogens in support of legislative and policy —efforts in food security.”

Tip: When citing pyani in your text, please consider quoting the precise release version (if you are using a specific release), or the specific commit hash (if you are using the ‘bleeding-edge’ development version), e.g. 

"...using PYANI (v0.3.0)...

or

"...using PYANI (commit bb1cb5c)...

This will enable precise reproduction of this part of your work by others.

5.1.1 Publications citing pyani

If you are using pyani, you are in good company. These authors and manuscripts have employed pyani to help with their whole-genome classification:

(If I’ve missed out your paper, please do put in a pull request on GitHub or get in touch and I’ll be happy to add it!)

2020


• Ibarra Caballero et al. (2020) “Genome comparison and transcriptome analysis of the invasive brown root rot pathogen, Phellinus noxius, from different geographic regions reveals potential enzymes associated with degradation of different wood substrates” Fungal Biology doi:10.1016/j.funbio.2019.12.007


• Kim et al. (2020) “Genome analysis of Lactobacillus plantarum subsp. plantarum KCCP11226 reveals a well-conserved C30 carotenoid biosynthetic pathway” 3 Biotech. doi:10.1007/s13205-020-2149-y


• Miyoshi et al. (2020) “Early-life microbial intervention reduces colitis risk promoted by antibiotic-induced gut dysbiosis” bioRxiv doi:10.1101/2020.03.11.987412

• Mu et al. (2020) “Bradymonabacteria, a novel bacterial predator with versatile survival strategies in saline environments” Microbiome doi:10.21203/rs.2.20535/v1


• Norris et al. (2020) “Acidithiobacillus ferriusus sp. nov.: an ancestral extremely acidophilic and facultatively anaerobic chemolithoautotroph” Extremophiles doi:10.1007/s00792-020-01157-1


• Pasanen et al. (2020) “Pectobacterium parvum sp. nov., having a Salmonella SPI-1-like Type III secretion system and low virulence” Int. J. Syst. Evol. Microb. doi:10.1099/ijsem.0.004057

• Rackaityte (2020) “Viable bacterial colonization is highly limited in the human intestine in utero” Nature Medicine doi:10.1038/s41591-020-0761-3


• Vijayan et al. (2020) “Bacteria known to induce settlement of larvae of Hydroides elegans are rare in natural inductive biofilm” Aquatic Microb. Ecol. doi:10.3354/ame01925

5.1. Citing pyani

• Wang et al. (2020) “Comparative Genomics Analysis of Lactobacillus ruminis from Different Niches” Genes doi:10.3390/genes11010070


• Wist et al. (2020) “Phenotypic and Genotypic Traits of Vancomycin-Resistant Enterococci from Healthy Food-Producing Animals” Microorganisms doi:10.3390/microorganisms8020261


2019


• Acevedo et al. (2019) “Bacillus clarus sp. nov. is a new Bacillus cereus group species isolated from soil” BioRxiv doi:10.1101/508077


• Boukerb et al. (2019) “Campylobacter armoricus sp. nov., a novel member of the Campylobacter lari group isolated from surface water and stools from humans with enteric infection” Int. J. Syst. Evol. Micro. doi:10.1099/ijsem.0.003836


• Doud et al. (2019) “Function-driven single-cell genomics uncovers cellulose-degrading bacteria from the rare biosphere” ISME J. doi:10.1038/s41396-019-0557-y


• Feng et al. (2019) “Complete genome sequence of Hahella sp. KA22, a prodigiosin-producing algicidal bacterium” Marine Genomics doi:10.1016/j.margen.2019.04.003


• Ghosh et al. (2019) “Reanalysis of Lactobacillus paracasei Lbs2 Strain and Large-Scale Comparative Genomics Places Many Strains into Their Correct Taxonomic Position” Microorganisms doi:10.3390/microorganisms7110487


• Kaminsky et al. (2019) “Genomic Analysis of γ-Hexachlorocyclohexane-Degrading Sphingopyxis lindanitolerans W55A3p Strain in the Context of the Pangenome of Sphingopyxis” Genes doi:0.3390/genes10090688

• Khan et al. (2019) “Genomic and physiological analyses reveal that extremely thermophilic Caldicellulosiruptor changbaiensis deploys unique cellulose attachment mechanisms” BioRxiv doi:10.1101/622977


• Kiu et al. (2019) “Phylogenomic analysis of gastroenteritis-associated Clostridium perfringens in England and Wales over a 7-year period indicates distribution of clonal toxigenic strains in multiple outbreaks and extensive involvement of enterotoxin-encoding (CPE) plasmids" Micro. Genom. doi:10.1099/mgen.0.000297


• Kochetkova et al. (2019) “Tepidiforma bonchosmolovskayae gen. nov., sp. nov., a moderately thermophilic Chloroflexi bacterium from a Chukotka hot spring (Arctic, Russia), representing a novel class, Tepidiformia, which includes the previously uncultivated lineage OLB14” Int. J. Syst. Evol. Microbiol. doi:10.1099/ijsem.0.003902


• Nordmann et al. (2019) “Complete genome sequence of the virus isolate vB_BthM-Goe5 infecting Bacillus thuringiensis” Arch. Virol. doi:10.1007/s00705-019-04187-z


• Stevens et al. (2019) “Whole-genome-based phylogeny of Bacillus cytotoxicus reveals different clades within the species and provides clues on ecology and evolution” Sci. Rep. doi:10.1038/s41598-018-36254-x


• Thorell et al. (2019) “Isolates from colonic spirochaetosis in humans show high genomic divergence and carry potential pathogenic features but are not detected by 16S amplicon sequencing using standard primers for the human microbiota” BioRxiv doi:10.1101/544502


• Vincent et al. (2019) “A Mesophilic Aeromonas salmonicida Strain Isolated from an Unsuspected Host, the Migratory Bird Pied Avocet” Microorganisms doi:10.3390/microorganisms7120592


• Zakham et al. (2019) “Molecular diagnosis and enrichment culture identified a septic pseudoarthrosis due to an infection with Erysipelatoclostridium ramosum” Int. J. Inf. Dis. doi:10.1016/j.ijid.2019.02.001


2018


• Ferretti et al. (2018) “Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome” Cell Host Microbe doi:10.1016/j.chom.2018.06.005


• Hubbard et al. (2018) “Comparison of the first whole genome sequence of *Haemophilus quentini* with two new strains of *Haemophilus quentini* and other species of *Haemophilus*** Genome doi:10.1139/gen-2017-0195


• Jung et al. (2018) “Complete genome sequence of *Bifidobacterium choerinum* FMB-1, a resistant starch-degrading bacterium” *J. Biotech.* doi:10.1016/j.jbiotec.2018.03.009


• McCann et al. (2018) “Viromes of one year old infants reveal the impact of birth mode on microbiome diversity” *PeerJ* doi:10.7717/peerj.4694


• Samad et al. (2017) “Comparative genome analysis of the vineyard weed endophyte *Pseudomonas viridiflava* CDRTe14 showing selective herbicidal activity” *Sci. Rep.* doi:10.1038/s41598-017-16495-y

• Sant’Anna et al. (2018) “Genome-based reclassification of *Paenibacillus dauci* as a later heterotypic synonym of *Paenibacillus shenyangensis*** Int. J. Syst. Evol. Micro. doi:10.1099/ijsem.0.003127


5.1. Citing pyani

• Vincent & Charette (2018) “Completion of genome of Aeromonas salmonicida subsp. salmonica 01-B526 reveals how sequencing technologies can influence sequence quality and result interpretations” New Microb. New Inf. doi:10.1016/j.nnni.2018.02.007

• Wilhelm (2018) “Following the terrestrial tracks of Caulobacter - redefining the ecology of a presumed aquatic oligotroph” ISME J doi:10.1038/s41396-018-0257-z


2017

• Anderson et al. (2017) “Genomic variation in microbial populations inhabiting the marine subseafloor at deep-sea hydrothermal vents” Nat. Comm. doi:10.1038/s41467-017-01228-6

• Ding et al. (2017) “Loss of the ssrA genome island led to partial debromination in the PBDE respiring Dehalococcoides mccartyi strain GY50” Env. Micro. doi:10.1111/1462-2920.13817


• Tohno et al. (2017) “Lactobacillus silagincola sp. nov. and Lactobacillus pentosiphilus sp. nov., isolated from silage” Int. J Syst. Evol. Biol doi:10.1099/ijsem.0.002196


2016

• Burstein et al. (2016) “New CRISPR–Cas systems from uncultivated microbes” Nature doi:10.1038/nature21059
• Tanizawa et al. (2016) “DFAST and DAGA: web-based integrated genome annotation tools and resources” Biosci. Microb. Food Health doi:10.12938/bmfh.16-003

5.2 About pyani

pyani is a Python package and standalone program for calculation of whole-genome similarity measures. It is designed to be used with draft or complete prokaryote (bacterial and archaeal) genomes, and implements the following methods:

• ANIb (average nucleotide identity using BLAST+)
• ANIblastall (average nucleotide identity using legacy BLAST)
• ANIm (average nucleotide identity using MUMmer)
• TETRA (4-mer sequence profiles)

5.2.1 Funding and Support

The development of pyani has been made possible by generous funding from a number of sources, including:

• The Scottish Government
  – 2011-2016 workpackages
  – 2016-2019 workpackage RD2.1.4
As is the case with much bioinformatics software, ongoing development does not receive continuous direct funding support. If you use pyani, please do cite this tool in your publications and outputs - this helps us justify future funding applications to maintain and improve the software.

You can find more information about how to cite pyani on the Citing pyani page.

## 5.3 QuickStart Guide

### 5.3.1 Installation

To use pyani you will need to install it on a local machine (your laptop, desktop, server or cluster). Installation is easiest using one of the two more popular Python package managers:

1. **conda**

   pyani is available through the bioconda channel of Anaconda:
   ```
   conda install -c bioconda pyani
   ```

2. **PyPI**

   pyani is available via the PyPI package manager for Python:
   ```
   pip install pyani
   ```

   **Tip:** pyani can also be installed directly from source, or run as a Docker image. More detailed, and alternative, installation instructions can be found on the Installation Guide page.

### 5.3.2 pyani Walkthrough

The general procedure for any pyani analysis is:

1. Collect genomes for analysis (and index them)
2. Create a database to hold genome data and analysis results
3. Perform ANI/TETRA/etc. analysis
4. Report and visualise analysis results
5. Generate species hypotheses (classify the input genomes) using the analysis results

To see options available for the pyani program, use the \(-h\) (help) option:

```pyani -h```

An example ANIm (ANI with MUMmer) analysis using pyani is provided as a walkthrough below. The sequence of commands used is:
Tip: If you have the pyani source code, you can run the walkthrough commands by executing make walkthrough at the command-line, in the repository root. You can clean up the walkthrough output with make clean_walkthrough.

1. Collect genomes

It is possible to use genomes you have already placed into a local directory with pyani, but for this walkthrough a new set of genomes will be obtained from GenBank, using the pyani download command.

Tip: To read more about using local files with pyani, please see the Indexing Genomes documentation. To read more about downloading genomes from NCBI, please see the Downloading Genomes from NCBI documentation.

Attention: To use their online resources programmatically, NCBI require that you provide your email address for contact purposes if jobs go wrong, and for their own usage statistics. This should be specified with the --email <EMAIL ADDRESS> argument of pyani download.

Using the pyani download subcommand, we download all available genomes for Candidatus Blochmannia from NCBI. The taxon ID for this grouping is 203804, and this ID is passed as the -t argument. The final (compulsory) argument is the path to the directory into which the genome data will be downloaded.

This creates a new directory (C_blochmannia) with the following contents:

```
$ tree C_blochmannia
C_blochmannia
    GCF_000011745.1_ASM1174v1_genomic.fna
    GCF_000011745.1_ASM1174v1_genomic.fna.gz
    GCF_000011745.1_ASM1174v1_genomic.md5
    ...
    GCF_000973545.1_ASM97354v1_hashes.txt
    classes.txt
    labels.txt
```

Each downloaded genome is represented by four files: the genome sequence (FASTA: *.fna, compressed: *.fna.gz), an NCBI hashes file (*.hashes.txt) and an MD5 hash of the genome sequence file (*.md5).

Two additional files are created, summarising all genomes in the subdirectory:
- classes.txt: defines a class to which each input genome belongs. This is used for determining membership of groups for each genome, and annotating graphical output.

5.3. QuickStart Guide
2. Create database

pyani uses a local SQLite3 database to store genome data and analysis results. Existing databases can be re-used. For this walkthrough we create a new, empty database by executing the command:

```bash
pyani createdb
```

**Tip:** This creates the new database in a default location (`.pyani/pyanidb`), but the name and location of this database can be controlled with the `pyani createdb` command (see the [Creating a Local pyani Database](#) documentation). The path to the database can be specified in each of the subsequent commands, to enable maintenance and sharing of multiple analysis runs.

3. Conduct ANIm analysis

We run ANIm on the downloaded genomes by specifying first the directory containing the genome data (here, `C_blochmannia`) then the path to a directory which will contain the analysis results (`C_blochmannia_ANIm` for this walkthrough).

We also provide a name for the analysis (`--name`, for later human-readable reference), with optional files defining labels for each genome to be used when plotting output (`--labels`) and a set of classes to which each genome belongs (`--classes`) for downstream analysis:

```bash
pyani anim C_blochmannia C_blochmannia_ANIm
   --name "C. blochmannia run 1"
   --labels C_blochmannia/labels.txt --classes C_blochmannia/classes.txt
```

This command runs ANIm analysis on the genomes in the specified `C_blochmannia` directory. As we did not specify a database, the analysis results will be stored in the default database we created earlier (`.pyani/pyanidb`), where they will be identified by the name `C. blochmannia run 1`. The comparison result files will be written to the `C_blochmannia_ANIm` directory.

4. Reporting Analyses and Analysis Results

We can list all the runs contained in the (default) database, using the command:

```bash
pyani report --runs C_blochmannia_ANIm/ --formats html,excel,stdout
```

This will report the relevant information to new files in the `C_blochmannia_ANIm` directory.

```
$ tree -L 1 C_blochmannia_ANIm/
C_blochmannia_ANIm/
    nucmer_output
    runs.html
    runs.tab
    runs.xlsx
```

**Tip:** By default the `pyani report` command will create a tab-separated text file with the `.tab` suffix, but by using the `--formats` option, we have also created an HTML file, and an Excel file with the same data. The `stdout` option
also prints the output table to the terminal window.

By inspecting the `runs.tab` file (or any of the other `runs.*` files) we see that our walkthrough analysis has run ID 1. So we can use this ID to get tables of specific information for that run, such as:

**the genomes that were analysed in all runs**

```
pyani report --runs_genomes --formats html,excel,stdout C_blochmannia_ANIm/
```

**the complete set of pairwise comparison results for a single run** (listed by comparison)

```
pyani report --run_results 1 --formats html,excel,stdout C_blochmannia_ANIm/
```

**comparison results as matrices** (percentage identity and coverage, number of aligned bases and “similarity errors”, and a Hadamard matrix of identity multiplied by coverage).

```
pyani report --run_matrices 1 --formats html,excel,stdout C_blochmannia_ANIm/
```

**Attention:** The `--run_results` and `--run_matrices` options take a single run ID or a comma-separated list of IDs (such as `1,3,4,5,9`) as an argument, and will produce output for each specified run ID.

**Graphical output**

Graphical output is obtained by executing the `pyani plot` subcommand, specifying the output directory and run ID. Optionally, output file formats and the graphics drawing method can be specified.

```
pyani plot --formats png,pdf --method seaborn C_blochmannia_ANIm 1
```

Supported output methods are:

- seaborn
- mpl (matplotlib)
- plotly

and each generates five plots corresponding to the matrices that `pyani report` produces:

- percentage identity of aligned regions
- percentage coverage of each genome by aligned regions
- number of aligned bases on each genome
- number of “similarity errors” on each genome
- a Hadamard matrix of percentage identity multiplied by percentage coverage for each comparison

Several graphics output formats are available, including `.png`, `.pdf` and `.svg`.

### 5.4 Requirements

The `pyani` package requires several other programs, packages and tools to run and develop. Many of these are automatically installed alongside `pyani`, but some packages and tools must be installed separately.

This page describes requirements for `pyani` and how/why they are used.
Fig. 1: Percentage identity matrix for *Candidatus* Blochmannia ANIm analysis

Each cell represents a pairwise comparison between the named genomes on rows and columns, and the number in the cell is the pairwise identity of aligned regions. The dendrograms are single-linkage clustering trees generated from the matrix of pairwise identity results. The default colour scheme colours cells with identity > 0.95 as red, and those with < 0.95 as blue. This division corresponds to a widely-used convention for bacterial species boundaries.
Fig. 2: Percentage coverage matrix for *Candidatus Blochmannia* ANIm analysis

Each cell represents a pairwise comparison between the named genomes on rows and columns, and the number in the cell is pairwise coverage of each genome by aligned regions in the comparison. The dendrograms are single-linkage clustering trees generated from the matrix of pairwise coverage results. The default colour scheme colours cells with identity > 0.50 as red, and those with < 0.50 as blue. This division corresponds to a strict majority of each genome in the comparison being alignable (a plausible minimum requirement for two sequences being considered “the same thing”).

5.4. Requirements
5.4.1 Python3

**pyani** is written in **Python**, and the modern version of **Python** is **Python3**. The legacy version of **Python** will not be maintained past 2020. **pyani** is written to use many features of **Python3** and will not run on **Python2**.

- **Python3**

5.4.2 NCBI-BLAST+

To carry out ANIb (average nucleotide identity using **BLAST**) analysis, genome sequences are compared using the **BLAST+** tool, provided by NCBI. The **BLAST+** tool is the current, maintained version, and is completely rewritten with respect to the legacy **BLAST** package (see below).

- **NCBI-BLAST+**

5.4.3 MUMmer v3.23

To carry out ANIm (average nucleotide identity using **MUMmer**) analysis, genome sequences are compared using the **nucmer** tool, part of the **MUMmer** package. Currently, **pyani** uses an older version of **MUMmer** for this analysis, pinned at version 3.23. **pyani** has not yet been tested with **MUMmer 4.x**.

- **MUMmer3**
- **MUMmer4**

5.4.4 Legacy NCBI-BLAST

An alternative implementation of ANIb (average nucleotide identity using **BLAST**), included for compatibility checks with other ANI calculation software is provided in **pyani** through the legacy script `average_nucleotide_identity.py`. The use of the legacy `aniblastall` analysis is not recommended, and NCBI do not recommend use of the legacy **NCBI-BLAST** tool. However, the legacy software can still be downloaded and installed, for the curious and those who wish to test legacy compatibility.

- **Legacy NCBI-BLAST (not supported)**

5.4.5 SQLite3

The output generated by **pyani** analyses is stored in a local database, provided by **SQLite3**, for rapid querying and recovery. This allows for persistent storage of results without the need to keep the original alignment files, and for incremental addition of new analyses. **SQLite** is installed with **Python**

- **SQLite**

5.4.6 Open Grid Scheduler

When running on a cluster, **pyani** currently schedules jobs using the Sun Grid Engine/Open Grid Engine/Open Grid Scheduler syntax. Your cluster will require a compatible scheduler for **pyani** to distribute jobs appropriately:
• Open Grid Scheduler

5.4.7 Python Packages

pyani relies on functionality provided by a number of additional Python packages, and we gratefully acknowledge their contribution:

• Biopython: for working with biological data formats
• Matplotlib: for graphical output
• NetworkX: for graph calculations and representation
• Numpy: for matrix calculations
• OpenPyXL: for MicroSoft Excel output compatibility
• Pandas: for dataframe operations
• Pillow: for graphics manipulation and rendering
• SciPy: for scientific computing operations
• Seaborn: for graphical output
• SQLAlchemy: (pinned at v1.2.18 for compatibility reasons) for interaction with SQLite3
• tqdm: provides progress bars for user interaction

Development

We rely on a number of additional packages to aid pyani development, and if you set up a development environment as recommended in Contributing to pyani, then the following Python packages will be installed or expected to be present:

• bandit: to check for security issues in the codebase
• black: to enforce consistent, opinionated code formatting
• codecov: to generate code coverage output for the codecov.io service
• coverage: to generate code coverage output for local inspection
• doc8: to check docstring formatting syntax
• flake8: for code linting
• jinja2: for output/docfile templating
• pre-commit: for checking code style and quality prior to git commit
• pylint: for code linting
• pytest: to manage and run automated testing
• pytest-cov: to integrate pytest with codecov and coverage
• pytest-ordering: to ensure pytest test ordering
• sphinx: to generate documentation
• sphinx-rtd-theme: to provide local ReadTheDocs style formatting
5.5 Installation Guide

We support four ways to install and run pyani on your system:

1. Installation with Anaconda (i.e. the conda package manager) [Recommended]
2. Installation via pip (i.e. from PyPI)
3. Installation from source (i.e. download from GitHub)
4. Installation of a Docker image

Note: If you wish to contribute to development of pyani, you will require developer tools and packages that are not included in these installation instructions. To set up a local environment suitable for developing pyani, please refer to the Contributing to pyani page.

5.5.1 1. Installation with Anaconda

Anaconda <https://www.anaconda.com/> is a Python language distribution that includes the conda package manager. Several channels, themed collections of packages, are available through the conda package manager. The latest release of pyani should always be available from the bioconda channel.

We recommend installation of pyani using the Anaconda3 or Miniconda3 distributions, and the conda package manager. This provides a straightforward way of managing third-party tool (e.g. MUMmer and NCBI-BLAST) installation, and allows for installation of pyani in a virtual environment, protected from other installations on the system.

1.1 Install Anaconda or Miniconda

If not already available on your system, install Anaconda3 or Miniconda3 on your system, following the instructions for your system on their respective websites:

- Anaconda3 download
- Miniconda3 download

After installation, you can check whether conda is available by issuing the following command in your terminal:

```
$ conda -V
conda 4.7.10
```

1.2 (optional) Create a new conda environment

Creation of a new conda environment is not necessary for installation of pyani, but it can be sometimes useful to specify the available version of Python, or to keep the versions of third party tools and Python packages separate from the system-level install. This may be particularly useful if, for example, the default system-level Python is Python2, as Python3 is required for pyani.

The following commands will create and activate a new conda environment called pyani, using Python 3.7:

```
conda create --name pyani --yes python=3.7  # --yes accepts all suggestions
conda activate pyani
```
If successful, you should see the prefix (pyani) before your terminal prompt. To exit the current conda environment, issue the following command:

```bash
conda deactivate
```

- Managing conda environments

### 1.3 Add required conda channels

pyani and many of the packages it requires are provided in the conda channel called bioconda. Channels are locations where conda will look for packages, and they are typically grouped by topic area.

To install pyani, you will need to make available the defaults, conda-forge and bioconda channels, with the following commands:

```bash
conda config --add channels defaults
conda config --add channels bioconda
conda config --add channels conda-forge
```

- Managing conda channels

### 1.4 Install pyani

The bioconda distribution of pyani will install all necessary packages and software required to run the software, including NCBI-BLAST+ and MUMmer, with the following command:

```bash
conda install --yes pyani
```

When installation is complete, you can check for the availability of the pyani programs with the following commands:

```bash
$ pyani --version
pyani 0.2.9
$ average_nucleotide_identity.py --version
average_nucleotide_identity.py: pyani 0.2.9
$ genbank_get_genomes_by_taxon.py --version
genbank_get_genomes_by_taxon.py: pyani 0.2.9
```

**Attention:** If you wish to use the ANIblastall legacy ANIb method, then the legacy NCBI-BLAST tools need to be installed. These are not available through conda and must be installed manually for your system, as described below.

### 5.5.2 2. Installation with pip

PyPI is the Python Packaging Index, a repository of software for the Python language. Packages from PyPI can be installed using the pip package installer, which should come preinstalled with your system's Python3. The latest release of pyani should always be available from PyPI.
2.1 (optional) Create and activate a new virtual environment

As with the conda installation route above, it can be useful to separate the version of Python you use for pyani, and any installed packages and tools, from the system-level Python installation. There are multiple tools available to do this, and for convenience we list some below:

- Anaconda distribution
- pipenv
- pyenv
- virtualenv

2.2 Install third-party tools

Two third-party software tools are needed to perform ANIm and ANIb analysis:

- MUMmer3 for ANIm
- NCBI-BLAST+ for ANIb

These tools are not part of the PyPI distribution of pyani, and should be installed according to the instructions on their respective websites.

Tip: If you are using a conda environment for pyani, you can install both tools with a single command: conda install --yes blast mummer.

Attention: If you wish to use the ANIblastall legacy ANIb method, then the legacy NCBI-BLAST tools need to be installed. These are not available through conda and must be installed manually for your system, as described below.

2.3 Install pyani

The pyani programs, and their Python dependencies, can be installed with the command:

```
pip install pyani
```

When installation is complete, you can check for the availability of the pyani programs with the following commands:

```
$ pyani --version
pyani 0.2.9
$ average_nucleotide_identity.py --version
average_nucleotide_identity.py: pyani 0.2.9
$ genbank_get_genomes_by_taxon.py --version
genbank_get_genomes_by_taxon.py: pyani 0.2.9
```

5.5.3 3. Installation from source

The source code for the current pyani release can always be found on GitHub:

- Current pyani release source
3.1 (optional) Create and activate a new virtual environment

As with the conda installation route above, it can be useful to separate the version of Python you use for pyani, and any installed packages and tools, from the system-level Python installation. There are multiple tools available to do this, and for convenience we list some below:

- Anaconda distribution
- pipenv
- pyenv
- virtualenv

3.2 Install third-party tools

Two third-party software tools are needed to perform ANIm and ANIb analysis:

- MUMmer3 for ANIm
- NCBI-BLAST+ for ANIb

These tools are not part of the PyPI distribution of pyani, and should be installed according to the instructions on their respective websites.

Tip: If you are using a conda environment for pyani, you can install both tools with a single command: conda install --yes blast mummer.

Attention: If you wish to use the ANIblastall legacy ANIb method, then the legacy NCBI-BLAST tools need to be installed. These are not available through conda and must be installed manually for your system, as described below.

3.3 Download and extract the pyani source code

Click on one of the two links on the pyani releases page, or use a download tool such as curl or wget to download the link to a convenient location. For example:

```shell
curl https://github.com/widdowquinn/pyani/archive/v0.2.9.tar.gz
```

Then extract the source code archive. This will create a new directory called pyani-<CURRENT_VERSION> in your current location:

```shell
$ tar -zxf v0.2.9.tar.gz
$ ls
[...]
pyani-0.2.9/
[...]
```

3.4 Install pyani

Change directory in the terminal to the source code directory, e.g.:
pyani can be installed using Python3’s setup tools, using the command:

```
python setup.py install
```

This will download and install the Python packages that pyani needs to run. When installation is complete, you can check for the availability of the pyani programs with the following commands:

```
$ pyani --version
pyani 0.2.9
$ average_nucleotide_identity.py --version
average_nucleotide_identity.py: pyani 0.2.9
$ genbank_get_genomes_by_taxon.py --version
genbank_get_genomes_by_taxon.py: pyani 0.2.9
```

### 5.5.4 4. Installation with Docker

Docker is a platform for running applications that uses containerisation to share virtual machines that come pre-installed with all required tools and dependencies. The latest pyani programs should always be available as separate Docker containers from DockerHub:

- `average_nucleotide_identity.py`
- `genbank_get_genomes_by_taxon.py`

In order to use the containerised versions of pyani, you must have Docker installed and working on your system. To do so, please follow the instructions at the Docker website:

- Docker website

#### 4.1 Running pyani with Docker

To pull (if necessary) and run the pyani programs in a Docker image on your local system, use the following commands (with the Docker daemon running):

```
docker run -v ${PWD}:/host_dir leightonpritchard/average_nucleotide_identity:v<REQUIRED_VERSION>
docker run -v ${PWD}:/host_dir leightonpritchard/genbank_get_genomes_by_taxon:v<REQUIRED_VERSION>
docker run -v ${PWD}:/host_dir leightonpritchard/pyani:v<REQUIRED_VERSION>
```

**Tip:** If no tag is specified, then Docker will attempt to use the :latest tag, which may not exist.
Note: The \(-v \{\text{PWD}\}:/host\_dir\) links the Docker image to the current working directory, and enables the \textit{pyani} programs to see files below the current execution point in your filesystem. This is necessary for the analysis to proceed.

5.6 Basic Use

5.6.1 Downloading Genomes from NCBI

This page describes some typical use cases for downloading genomes from NCBI using the \textit{pyani download} subcommand. This command downloads all assembled genomes found beneath a given taxon identifier, at the NCBI GenBank database. You will need to know the NCBI taxon ID for each taxon you wish to download.

\textbf{Attention:} To use their online resources programmatically, NCBI require that you provide your email address for contact purposes if jobs go wrong, and for their own usage statistics. This should be specified with the \textit{--email} \texttt{<EMAIL ADDRESS>} argument of \textit{pyani download}.

For more information about the \textit{pyani download} subcommand, please see the \textit{pyani download} page, or issue the command \textit{pyani download -h} to see the inline help.

Download all genomes in a single taxon

The basic form of the command is:

\begin{verbatim}
pyani download --email my.email@my.domain -t <TAXON_ID> <OUTPUT_DIRECTORY>
\end{verbatim}

This instructs \textit{pyani} to use the \textit{download} subcommand to obtain all available genome assemblies below the taxon ID \texttt{<TAXON_ID>}, passed with the \texttt{-t} argument, and place the downloaded files - along with label and class information files created by \textit{pyani} in the subdirectory \texttt{<OUTPUT_DIRECTORY>}.

For example, if we wished to download all available assemblies for the bacterium \textit{Pseudomonas flexibilis} we would identify the taxon ID to be 706570, and use this as the argument to \texttt{-t}, placing the output in a convenient subdirectory (e.g. \texttt{genomes}), with the command:

\begin{verbatim}
$ pyani download --email my.email@my.domain -t 706570 genomes
\end{verbatim}

This displays each assembly as a download is attempted, and places all output in the named subdirectory:
Each genome is downloaded in compressed format (.fna.gz files) and expanded in-place to give the FASTA file (.fna) files. The MD5 hash of each FASTA file is also calculated (.md5). This will be used by pyani to uniquely identify that assembly throughout the analysis process.

```
$ head genomes/GCA_001312105.1_ASM131210v1_genomic.md5
e55cd3d913a198ac60afd8d509c02ab4 genomes/GCA_001312105.1_ASM131210v1_genomic.fna
```

pyani also creates two files:

- **classes.txt**: each genome is assigned a class which is used to annotate genomes in the graphical output. pyani attempts to infer genus and species as the default class.

- **labels.txt**: each genome is assigned a text label, which is used to label genomes in the graphical output. pyani attempts to infer genus, species, and strain ID as the default label.

```
$ head genomes/classes.txt
1ac4941c569f32f044eba0a8540d4704 GCF_900155995.1_IMG-taxon_2596583557_annotated_genomic
  Pseudomonas flexibilis
8664341798070f8d70b2596a5b3a2320 GCF_900101515.1_IMG-taxon_2596583557_annotated_genomic
  Pseudomonas flexibilis
34 Chapter 5. Reporting problems and requesting improvements
```

(continues on next page)
These files are used to associate labels and classes to the genome files in the pyani database, specific to the analysis run.

Both classes.txt and labels.txt can be edited to suit the user’s classification and labelling scheme.

### Download all genomes from multiple taxa

To download genomes from more than one taxon, you can provide a comma-separated list of taxon IDs to the **pyani** `download` subcommand, e.g.:

```
pyani download --email my.email@my.domain -t <TAXON_ID1>,<TAXON_ID2>,... <OUTPUT_DIRECTORY>
```

The following command can be used to download assemblies from three different *Pseudomonas* taxa (*P. flexibilis*: 706570, *P. mosselli*: 78327, and *P. fulva*: 47880):

```
$ pyani download --email my.email@my.domain -t 706570,78327,47880 multi_taxa
```

**Dry-run test (identify, but do not download, files)**

If you only want to see which genomes will be downloaded from NCBI with a given **pyani** `download` subcommand, but not download them, then you can use the `--dry-run` option. For example:

```
$ pyani download --email my.email@my.domain -t 706570,78327,47880 multi_taxa --dry-run
```

**Download genomes for compilation of a custom Kraken database**

Kraken is a bioinformatics tool that assigns taxonomic identities to short DNA sequences, such as Illumina or Nanopore reads. Several wide-ranging Kraken databases are available for download, and around the community, but it can sometimes be useful to construct a custom local Kraken database specific for your organism or taxon of interest (e.g. for filtering out contaminating or suspect reads).

The **pyani** `download` command can prepare downloaded genome files for immediate use with the Kraken database-building tools, by specifying the `--kraken` option:
$ pyani download --email my.email@my.domain -t 706570,78327,47880 genomes_kraken --kraken
GCF_900155995.1_IMG-taxon_2681812811_annotated_assembly_genomic.fna.gz: 2097152it → [00:00, 3085741.03it/s]
GCF_900155995.1_IMG-taxon_2681812811_annotated_assembly_hashes.txt: 1048576it [00:00, → 140958511.30it/s]
WARNING: Modifying downloaded sequence for Kraken compatibility
GCF_900101515.1_IMG-taxon_2596583557_annotated_assembly_genomic.fna.gz: 2097152it → [00:01, 1902023.68it/s]
GCF_900101515.1_IMG-taxon_2596583557_annotated_assembly_hashes.txt: 1048576it [00:00, → 31428985.47it/s]
WARNING: Modifying downloaded sequence for Kraken compatibility

Using this option does affects downstream performance or use of pyani only in that the two different input files for the same genome will have distinct hashes:

$ head multi_taxa/GCA_001312105.1_ASM131210v1_genomic.fna
>BBCY01000001.1 Pseudomonas tuomuerensis JCM 14085 DNA, contig: JCM14085.contig00001, → whole genome shotgun sequence
ACCAGCATCTGGCGGATCAGGTCGCGGGCCTTCTCGGCCGATTGGCGGATGCGCCCGAGGTAGCGGCCGAGCGGCGCGTCGCCGCGCTCGCCCGCCAGCTCCTCGGCCATCTGCGTGTAGCCGAGCATGCTGGTCAGCAGGTTGTTGAAGTCGTGGGCAATGCCGCCGGTCAGGTGGCCGATGGCTTCCATGCGCTGCGCCTGGCGCAGCTGCTGTTCCAGCGCCGCCCGCTCCACCTCG

and so will not be considered to be the “same sequence” when repeating comparisons.

### 5.6.2 Indexing Genomes

Indexing genomes is a necessary step in pyani to prepare input genomes for analysis.

**Attention:** If you use the `pyani download` subcommand (see Downloading Genomes from NCBI) to obtain genomes for analysis, then indexing is carried out automatically. However, if you collect a local set of genomes (e.g., from your own sequencing project), then you will need to index the genomes with the `pyani index` subcommand.

For more information about the `pyani index` subcommand, please see the `pyani index` page, or issue the command `pyani index -h` to see the inline help.

**What does indexing do?**

In the context of pyani, *indexing* refers to generating an index code that is unique to each input genome FASTA file in the input directory. The index code is the MD5 hash for the FASTA file.
This MD5 index code is used to identify each specific input genome sequence (and associated metadata) so that duplicate comparisons can be readily identified, and previous results reused from the pyani database, if they are available.

Indexing also generates two files (see Downloading Genomes from NCBI):

- **classes.txt**: each genome is assigned a class which is used to annotate genomes in the graphical output. pyani attempts to infer genus and species as the default class
- **labels.txt**: each genome is assigned a text label, which is used to label genomes in the graphical output. pyani attempts to infer genus, species, and strain ID as the default label

These files are used to associate labels and classes to the genome files in the pyani database, specific to the analysis run. Both classes.txt and labels.txt can be edited to suit the user’s classification and labelling scheme.

### Index a directory of FASTA files

The basic form of the command is:

```
pyani index <GENOME_DIRECTORY>
```

This instructs pyani to search <GENOME_DIRECTORY> for files with a standard FASTA suffix (.fna, .fasta, .fa, .fas, .fsa_nt). For each file found, it calculates the MD5 hash and writes it to an accompanying file with extension .md5. The hash is then associated with a genome label and a genome class, written to the two files labels.txt and classes.txt (see above).

For example, if we have a directory called unindexed that contains some FASTA format genome sequence files:

```
$ tree unindexed
unindexed
    GCA_001312105.1_ASM131210v1_genomic.fna
    GCF_000834555.1_ASM83455v1_genomic.fna
    GCF_005796105.1_ASM579610v1_genomic.fna
```

We could run the pyani index command:

```
$ pyani index unindexed/
$ tree unindexed
unindexed
    GCA_001312105.1_ASM131210v1_genomic.fna
    GCA_001312105.1_ASM131210v1_genomic.md5
    GCF_000834555.1_ASM83455v1_genomic.fna
    GCF_000834555.1_ASM83455v1_genomic.md5
    GCF_005796105.1_ASM579610v1_genomic.fna
    GCF_005796105.1_ASM579610v1_genomic.md5
    classes.txt
    labels.txt
```

This creates an .md5 file for each genome, and corresponding classes.txt and labels.txt files:

```
$ head unindexed/GCA_001312105.1_ASM131210v1_genomic.fna
>BBCY01000001.1 Pseudomonas tuomerensis JCM 14085 DNA, contig: JCM14085.contig00001, ...
ACCAGCATCTGGCGGATCAGGTCGCGGGCCTTCTCGGCCGATTGGCGGATGCGCCCGAGGTAGCGGCCGAGCGGCGCGTC
GCGCGCGTCCGGCCGCGCAGCTCTCCTCGCCGTATGCGGCAATGCGGCAGGCGTGCGTAGCGAGCCGAGCGCGCGCGTC
GCCGCGCGTCGCGCGCGCGCCGCACGCCTCCTCGCCGTATGCGGCAATGCGGCAGGCGTGCGTAGCGAGCCGAGCGCGCGCGTC

$ head unindexed/GCA_001312105.1_ASM131210v1_genomic.md5
e55cd3d913a198ac60af8d509c02ab4 unindexed/GCA_001312105.1_ASM131210v1_genomic.fna

$ head unindexed/classes.txt
```

(continues on next page)
5.6.3 Creating a Local `pyani` Database

`pyani` stores genome information and analysis results in a persistent local SQLite3 database. This allows for reuse of previous comparisons and reanalysis of datasets without having to rerun the analysis. It also means that the genome comparison results don’t have to be stored in full on disk, saving space.

**Note:** To conduct a `pyani` analysis, there needs to be an existing database in-place.

To create a new, empty database you can use the `pyani createdb` command.

For more information about the `pyani createdb` subcommand, please see the `pyani createdb` page, or issue the command `pyani download -h` to see the inline help.

### Create a new empty `pyani` database

The basic form of the command is:

```
pyani createdb
```

This instructs `pyani` to create a new, empty database for analysis at the default location.

**Note:** The default location for the `pyani` database is in a hidden directory: `.pyani/pyanidb`. All other `pyani` subcommands will look in this location for the database, unless told otherwise using the `--dbpath` option.

For example:
Create an empty pyani database at a specific location

**Tip:** If you use pyani for a number of distinct taxa, it can be convenient to create a new database for each project, to avoid performance issues as the database grows in size, filled by data that does not contribute to the analysis.

The following command can be used to specify the location of the newly-created pyani database:

```
pyani createdb --dbpath <PATH_TO_DATABASE>
```

where `<PATH_TO_DATABASE>` is the intended location of the database. For instance, to create a new database specific for an analysis we’ll call multitaxa, we could use the command:

```
$ ls .pyani
pyanidb
$ pyani createdb --dbpath .pyani/multitaxadb
$ ls .pyani
multitaxadb pyanidb
```

The new database can then be specified in other pyani subcommands, using the `--dbpath` option.

### 5.6.4 Running ANIm analysis

pyani implements average nucleotide identity analysis using MUMmer3 (ANIm) as defined in Richter & Rosselló-Móra (2009) ([doi:10.1073/pnas.0906412106](https://doi.org/10.1073/pnas.0906412106)). To run ANIm on a set of input genomes, use the pyani anim sub-command.

In brief, the analysis proceeds as follows for a set of input prokaryotic genomes:

1. MUMmer3 is used to perform pairwise comparisons between each possible pair of input genomes, to identify homologous (alignable) regions.
2. For each comparison, the alignment output is parsed, and the following values are calculated:
   - total number of aligned bases on each genome
   - fraction of each genome that is aligned (the coverage)
   - the proportion of all aligned regions that is identical in each genome (the ANI)
   - the number of unaligned or non-identical bases (the similarity errors)
   - the product of coverage and ANI

The output values are recorded in the pyani database.

**Note:** A single MUMmer comparison is performed between each pair of genomes. Input genomes are sorted into alphabetical order by filename, and the query sequence is the genome that occurs earliest in the list; the subject sequence is the genome that occurs latest in the list.
Tip: The MUMmer comparisons are embarrassingly parallel, and can be distributed across cores on an Open Grid Scheduler-compatible cluster, using the `--scheduler SGE` option.

Attention: `pyani anim` requires that a working copy of MUMmer3 is available. Please see Installation Guide for information about installing this package.

For more information about the `pyani anim` subcommand, please see the `pyani anim` page, or issue the command `pyani anim --h` to see the inline help.

**Perform ANIm analysis**

The basic form of the command is:

```
pyani anim <INPUT_DIRECTORY> <OUTPUT_DIRECTORY>
```

This instructs `pyani` to perform ANIm on the genome FASTA files in `<INPUT_DIRECTORY>`, and write any output files to `<OUTPUT_DIRECTORY>`. For example, the following command performs ANIm on genomes in the directory `genomes` and writes output to a new directory `genomes_ANIm`:

```
pyani anim genomes genomes_ANIm
```

**Note:** While running, `pyani anim` will show progress bars unless these are disabled with the option `--disable_tqdm`.

This command will write the intermediate nucmer/MUMmer output to the directory `genomes_ANIm`, in a subdirectory called `nucmer_output`, where the results can be inspected if required.

```
$ ls genomes_ANIm/
nucmer_output
```

Attention: To view the output ANIm results, you will need to use the `pyani report` or `pyani plot` subcommands. Please see `pyani report` and `pyani plot` for more details.

**Perform ANIm analysis with Open Grid Scheduler**

The MUMmer comparison step of ANIm is embarrassingly parallel, and nucmer jobs can be distributed across cores in a cluster using the Open Grid Scheduler. To enable this during the analysis, use the `--scheduler SGE` option:

```
pyani anim --scheduler SGE genomes genomes_ANIm
```

**Note:** Jobs are submitted as array jobs to keep the scheduler queue short.

**Note:** If `--scheduler SGE` is not specified, all MUMmer jobs are run locally with Python’s multiprocessing module.
Controlling parameters of Open Grid Scheduler

It is possible to control the following features of Open Grid Scheduler via the `pyani anim` subcommand:

- The array job size (by default, comparison jobs are batched in arrays of 10,000)
- The prefix string for the job, as reported in the scheduler queue
- Arguments to the `qsub` job submission command

These allow for useful control of job execution. For example, the command:

```
pyani anim --scheduler SGE --SGEgroupsize 5000 genomes genomes_ANIm
```

will batch MUMmer jobs in groups of 500 for the scheduler. The command:

```
pyani anim --scheduler SGE --jobprefix My_Ace_Job genomes genomes_ANIm
```

will prepend the string `My_Ace_Job` to your job in the scheduler queue. And the command:

```
pyani anim --scheduler SGE --SGEargs "-m e -M my.name@my.domain" 5000 genomes genomes_ANIm
```

will email `my.name@my.domain` when the jobs finish.

References


5.7 Examples

5.7.1 Using non-NCBI genomes

It is usual to want to include or work only with genomes that have been generated locally, or that were not downloaded from NCBI using `pyani download`. To use these genomes with the `pyani` analysis subcommands, the genomes must be indexed\(^1\).

To index a set of genomes, use the `pyani index` subcommand on the input directory. To index the directory `mygenomes`, for example:

```
pyani index mygenomes
```

This will create a `.md5` file (containing the hash) for each genome, as well as class and label files listing all the input genomes.

All genomes in the `mygenomes` directory will now be available for use in `pyani`.

**class and labels files**

Using the `pyani download` command will create two files, by default `classes.txt` and `labels.txt` containing identifiers for each input genome that are used in later analysis and visualisation. These files are also created when `pyani index` is used as above.

---

\(^1\) **indexing** here refers to constructing a hash of the genome: a short representation of the entire genome’s contents that can be used to identify it uniquely.
The location of the labels and classes files may be changed using the `--labels` and `--classes` arguments, for example:

```
pyani index mygenomes --classes myclasses.txt --labels mylabels.txt
```

## 5.8 pyani subcommands

`pyani` has a subcommand structure, where the command `pyani` is followed immediately by a subcommand to let the software know which action you wish it to perform, in the format `pyani <subcommand>`. For example, to create a database you would use the `createdb` subcommand:

```
pyani createdb
```

This document links out to detailed instructions for each of the `pyani` subcommands.

### 5.8.1 pyani download

The `download` subcommand controls download of genome files from the NCBI Assembly database for input to `pyani`.

```
                  --email EMAIL [--api_key API_KEYPATH]
                  [--retries RETRIES] [--batchsize BATCHSIZE]
                  [--timeout TIMEOUT] [-f] [--noclobber]
                  [--labels LABELFNAME] [--classes CLASSFNAME]
                  [--kraken] [--dry-run]

outdir
```

**Positional arguments**

`outdir` The `outdir` argument should be the path to a directory into which genome files will be downloaded. If the directory exists, a warning will be given and the download will not proceed, to avoid overwriting existing data. To force writing into an existing directory, use the `--force` option.

**Flagged arguments**

`--api_key PATH_TO_API_KEY` The program will attempt to use an NCBI API key (see here) located at `PATH_TO_API_KEY`. Default: `~/.ncbi/api_key`

`--batchsize BATCHSIZE` The download process will attempt to download assemblies in multiples of `BATCHSIZE`. Default: 10000

`--classes CLASSFNAME` Write a set of labels (one per downloaded genome) to the file `CLASSFNAME` in `outdir`. Default: `classes.txt`

`--disable_tqdm` Disable the `tqdm` progress bar while the download process runs. This is useful when testing to avoid aesthetic problems with test output.

`--dry-run` Perform all actions of the download process except for downloading files.

`--email EMAIL` COMPULSORY. Provide the email address `EMAIL` to NCBI so that they can track problems.

`-f, --force` Force use of the `outdir` directory when downloaded genomes, even if it already exists.
-h, --help  Display usage information for pyani download.

--kraken  Add taxonomy information to the FASTA file headers of downloaded genomes. This allows the genomes to be readily used to construct databases for the Kraken software package.

-l LOGFILE, --logfile LOGFILE  Provide the location LOGFILE to which a logfile of the download process will be written.

--labels LABELFNAME  Write a set of labels (one per downloaded genome) to the file LABELFNAME in outdir. Default: labels.txt

--noclobber  Do not overwrite individual files in the outdir directory, when used with -f.

--retries RETRIES  The download process will attempt to download each batch of assemblies a maximum of RETRIES times. Default: 20

-t TAXON, --taxon TAXON  COMPULSORY. All genomes below taxon ID TAXON of a node in the NCBI Taxonomy database will be downloaded to the location specified by outdir.

--timeout TIMEOUT  The download process will wait a amaximum of TIMEOUT seconds before abandoning a URL connection attempt. Default: 10

-v, --verbose  Provide verbose output to STDOUT

5.8.2 pyani index

The index subcommand will index the genome files it finds the passed directory indir, generating label and class files, and files that contain an MD5 hash of the nucleotide sequence of each genome.

```bash
usage: pyani index [-h] [-l LOGFILE] [-v] [--disable_tqdm]
                  [--labels LABELFNAME] [--classes CLASSFNAME]
                  indir
```

Positional arguments

indir  The indir argument should be the path to a directory containing genome sequence data as FASTA files (one per genome assembly).

Flagged arguments

--classes CLASSFNAME  Write a set of labels (one per genome sequence file) to the file CLASSFNAME in indir. Default: classes.txt

--disable_tqdm  Disable the tqdm progress bar while the download process runs. This is useful when testing to avoid aesthetic problems with test output.

-h, --help  Display usage information for pyani index.

-l LOGFILE, --logfile LOGFILE  Provide the location LOGFILE to which a logfile of the download process will be written.

--labels LABELFNAME  Write a set of labels (one per genome sequence file) to the file LABELFNAME in indir. Default: labels.txt

-v, --verbose  Provide verbose output to STDOUT

5.8. pyani subcommands
5.8.3 pyani createdb

The `createdb` subcommand creates a new, empty database for `pyani` to use in subsequent analysis runs.

```
                      [--dbpath DBPATH] [-f]
```

Flagged arguments

`--disable_tqdm` Disable the `tqdm` progress bar while the download process runs. This is useful when testing to avoid aesthetic problems with test output.

`--dbpath DBPATH` Path to the location where the database will be created. Default: `.pyani/pyanidb`

`-h, --help` Display usage information for `pyani index`.

`-l LOGFILE, --logfile LOGFILE` Provide the location `LOGFILE` to which a logfile of the download process will be written.

`-v, --verbose` Provide verbose output to `STDOUT`

5.8.4 pyani anim

The `anim` subcommand will carry out ANIm analysis using genome files contained in the `indir` directory, writing result files to the `outdir` directory, and recording data about each comparison and run in a local SQLite3 database.

```
                      [--scheduler {multiprocessing,SGE}] [--workers WORKERS]
                      [--SGEgroupsize SGEGROUPSIZE] [--SGEargs SGEARGS]
                      [--jobprefix JOBPREFIX] [--name NAME] [--classes CLASSES]
                      [--labels LABELS] [--recovery] [--dbpath DBPATH]
                      [--nucmer_exe NUCMER_EXE] [--filter_exe FILTER_EXE]
                      [--maxmatch] [--nofilter]
                      indir outdir
```

Positional arguments

`indir` Path to the directory containing indexed genome files to be used for the analysis.

`outdir` Path to a directory where comparison output files will be written.

Flagged arguments

`--classes CLASSFNAME` Use the set of classes (one per genome sequence file) found in the file `CLASSFNAME` in `indir`. Default: `classes.txt`

`--dbpath DBPATH` Path to the location of the local `pyani` database to be used. Default: `.pyani/pyanidb`

`--disable_tqdm` Disable the `tqdm` progress bar while the download process runs. This is useful when testing to avoid aesthetic problems with test output.

`--filter_exe FILTER_EXE` Path to the MUMmer delta-filter executable. Default: `delta-filter`

`-h, --help` Display usage information for `pyani index`. 
--jobprefix JOBPREFIX Use the string JOBPREFIX as a prefix for SGE job submission names. Default: PYANI

--labels LABELFNAME Use the set of labels (one per genome sequence file) found in the file LABELFNAME in indir. Default: labels.txt

--name NAME Use the string NAME to identify this ANIm run in the pyani database.

--nucmer_exe NUCMER_EXE Path to the MUMmer nucmer executable. Default: nucmer

-l LOGFILE, --logfile LOGFILE Provide the location LOGFILE to which a logfile of the download process will be written.

--maxmatch Use the MUMmer --maxmatch option to include all nucmer matches.

--nofilter Do not use delta-filter to restrict nucmer output to 1:1 matches.

--recovery Use existing NUCmer comparison output if available, e.g. if recovering from a failed job submission. Using this option will not generate a new comparison if the old output files exist.

--scheduler {multiprocessing, SGE} Specify the job scheduler to be used when parallelising genome comparisons: one of multiprocessing (use many cores on the current machine) or SGE (use an SGE or OGE job scheduler). Default: multiprocessing.

--SGEargs SGEARGS Pass additional arguments SGEARGS to qsub when running the SGE-distributed jobs.

--SGEGroupsize SGEGROUPSIZE Create SGE arrays containing SGEGROUPSIZE comparison jobs. Default: 10000

-v, --verbose Provide verbose output to STDOUT

--workers WORKERS Spawn WORKERS worker processes with the --scheduler multiprocessing option. Default: 0 (use all cores)

5.8.5 pyani anib

5.8.6 pyani report

5.8.7 pyani plot

5.8.8 pyani classify

5.8.9 pyani listdeps

The listdeps subcommand writes an account of the local platform, the installed Python version, and installed dependencies (Python packages and third-party software tools) to STDOUT.

usage: pyani.py listdeps [-h] [-l LOGFILE] [-v]

Flagged arguments

--disable_tqdm Disable the tqdm progress bar while the download process runs. Does nothing.

-h, --help Display usage information for pyani listdeps.

-l LOGFILE, --logfile LOGFILE Provide the location LOGFILE to which a logfile of the subcommand output will be written.

-v, --verbose Provide verbose output to STDOUT. This actually provides no additional information.

5.8. pyani subcommands
5.9 Testing

We are currently writing tests formatted for the pytest package, for testing pyani.

**Warning:** Some tests are still targeted at nosetests, which is in maintenance mode and, although we are still in transition, our plan is to change the test framework completely to use pytest.

5.9.1 Test directory structure

All tests, supporting input and target data, are to be found in the tests/ subdirectory of the pyani repository.

- Input data for tests is located under the tests/test_input directory, in subdirectories named for the general operation that is being tested.
- Target data (known correct values) for tests is located under the tests/test_targets directory, in subdirectories named for the general operation being tested.
- Test output is written to the tests/test_output directory, in subdirectories named for the general operation being tested.

**Tip:** If you wish to write new tests for pyani, we ask that your test data and operations conform to this structure.

The tests/test_failing_data directory contains data that is known to cause problems for pyani in that at least two of the input genomes have no appreciable sequence identity.

5.9.2 Running tests

To run tests with nosetests, change directory to the root of the pyani repository, and invoke a nosetests command.

**Run all tests locally**

To run all tests locally on your machine, issue the following command from the repository root:

```
nosetests -v
```

This will cause nose to run all tests under the tests/ subdirectory.

**Run individual tests**

Tests are grouped in files with filenames that match test_*.py. We aim to write tests as classes that subclass unittest.TestCase, as described in the nosetests documentation. An example of this style can be found in the tests/test_anim.py test file.

This style allows us to run tests at several levels of granularity, specifying all tests (see above), all tests within a module (e.g. test_anim.py), or all tests within a class in that test file.

For example, to run all ANIm-related tests, we can issue:

```
nosetests -v tests/test_anim.py
```
To run all tests of nucmer command line generation, we can specify a single class within that file using the command:

```bash
nosetests -v tests/test_anim.py:TestNUCmerCmdline
```

And to test only “multiple command generation” we can issue the following:

```bash
nosetests -v tests/test_anim.py:TestNUCmerCmdline.test_multi_cmd_generation
```

## 5.10 Contributing to pyani

### 5.10.1 Reporting bugs and errors

If you find a bug, or an error in the code or documentation, please report this by raising an issue at the GitHub issues page for pyani:

- [GitHub issues page](https://github.com/organization-name/pyani/issues)

### 5.10.2 Contributing code or documentation

We gratefully accept code and other contributions. A list of contributors can be found via the Github contributors link. You are welcome to help develop pyani, fix a bug, improve documentation, or contribute in any other way. To make everyone’s lives easier in this process, we ask that you please follow the guidelines for developers below:

**Pre-commit checks and style guide**

So far as is possible, we aim to follow the coding conventions as described in PEP8 and PEP257, but we have adopted black code styling, which does vary from the PEPs in places.

We use the `flake8` tool for style checks, and this can be installed as follows (with two useful plugins):

```bash
pip install flake8 flake8-docstrings flake8-blind-except
```

`flake8` can then be run directly on the codebase with

```bash
flake8 bin/
flake8_pyani/
```

We use the `black` tool for code style checking, which can be installed with:

```bash
pip install black
```

The `flake8` and `black` styles can be enforced as pre-commit hooks using the `pre-commit` package (included in `requirements.txt`).

The `black` and `flake8` hooks are defined in `.pre-commit-config.yaml`; custom settings for `flake8` are held in `.flake8` (all files are under version control).

To enable pre-commit checks in the codebase on your local machine (once `pre-commit` has been installed), execute the following command in the root directory of this repository:

```bash
pre-commit install
```
Checking changes to the documentation

Much of the repository documentation is written in Markdown files, but the main documentation (which you are reading) is prepared for ReadTheDocs, which uses reStructuredText and Sphinx. The Sphinx configuration is described in `docs/conf.py` (under version control).

So long as Sphinx is installed on your machine, you can check your documentation changes locally, building inplace by changing to the `docs/` directory and issuing:

```
make html
```

This will place a compiled version of the documentation under `_build/html`, which you can inspect before committing to the repository.

**Tip:** To build the documentation in ReadTheDocs style, you will need to install the corresponding theme with `pip install sphinx_rtd_theme` or `conda install sphinx_rtd_theme`

For now, docstrings in the source code are not required to be in any controlled syntax, such as reStructuredText, but this may change.

Making changes and pull requests

1. Fork the `pyani` repository under your account at GitHub.
2. Clone your fork to your development machine.
3. Create a new branch in your forked repository with an informative name like `fix_issue_107`, using `git` (e.g. with the command `git checkout -b fix_issue_107`).
4. Make the changes you need and commit them to your local branch.
5. Run the repository tests (see the Testing documentation for more details).
6. If the tests all pass, push the changes to your fork, and submit a pull request against the original repository.
7. Indicate one of the `pyani` developers as an assignee to review your pull request when you submit your pull request.

The assigned developer will then review your pull request, and merge it or continue the conversation, as appropriate.

5.10.3 Suggestions for improvement

If you would like to make a suggestion for how we could improve `pyani`, we welcome contributions at the GitHub issues page.

5.11 Licensing

Unless otherwise indicated, all code is subject to the following agreement:

(c) The James Hutton Institute 2014-2019 (c) The University of Strathclyde 2019 Author: Leighton Pritchard

Contact: leighton.pritchard@strath.ac.uk
Address: Leighton Pritchard, Strathclyde Institute of Pharmacy and Biomedical Sciences, University of Strathclyde, 161 Cathedral Street, Glasgow, G4 0RE, Scotland, UK

5.11.1 The MIT License

Copyright (c) 2014-2019 The James Hutton Institute Copyright (c) 2019 The University of Strathclyde

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the “Software”), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED “AS IS”, WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

5.12 pyani package

Module with main code for pyani application/package.

```
exception pyani.PyaniException
    Bases: Exception

    General exception for pyani.
```

5.12.1 Subpackages

pyani.scripts package

Modules to support use of pyani as a script.

Subpackages

pyani.scripts.parsers package

Module providing command-line parser definitions.

```
pyani.scripts.parsers.parse_cmdline(  
    argv: Optional[List[T]] = None) -> argparse.Namespace
```

Parse command-line arguments for script.

**Parameters**

- `argv` – Namespace, command-line arguments

The script offers a single main parser, with subcommands for the actions:

- **download** download all available NCBI assemblies below the passed taxonomy ID
- **index** index genome sequence files in a subdirectory, for analysis
• **createdb** generate SQLite database for data and analysis results
• **anim** conduct ANIm analysis
• **anib** conduct ANIb analysis
• **aniblastall** conduct ANIblastall analysis
• **report** generate output describing analyses, genomes, and results
• **plot** generate graphical output describing results
• **classify** produce graph-based classification of genomes on the basis of ANI analysis

**Submodules**

**pyani.scripts.parsers.anib_parser module**

Provides parser for anib subcommand.

```python
pyani.scripts.parsers.anib_parser.build(subps: argparse._SubParsersAction, parents: Optional[List[argparse.ArgumentParser]] = None) → None
```

Return a command-line parser for the anib subcommand.

**Parameters**

- **subps** – collection of subparsers in main parser
- **parents** – parsers from which arguments are inherited

The terminology may be confusing, but in practice the main parser collects command-line arguments that are then available to this parser, which inherits options from the parsers in **parents** in addition to those defined below.

**pyani.scripts.parsers.aniblastall_parser module**

Provides parser for aniblastall subcommand.

```python
pyani.scripts.parsers.aniblastall_parser.build(subps: argparse._SubParsersAction, parents: Optional[List[argparse.ArgumentParser]] = None) → None
```

Return a command-line parser for the aniblastall subcommand.

**Parameters**

- **subps** – collection of subparsers in main parser
- **parents** – parsers from which arguments are inherited

**pyani.scripts.parsers.anim_parser module**

Provides parser for anim subcommand.

```python
pyani.scripts.parsers.anim_parser.build(subps: argparse._SubParsersAction, parents: Optional[List[argparse.ArgumentParser]] = None) → None
```

Return a command-line parser for the anim subcommand.

**Parameters**
• **subps** – collection of subparsers in main parser
• **parents** – parsers from which arguments are inherited

**pyani.scripts.parsers.classify_parser module**

Provides parser for classify subcommand.

```python
def classify_parser.build(subps: argparse._SubParsersAction, parents: Optional[List[argparse.ArgumentParser]] = None) -> None:
    Return a command-line parser for the classify subcommand.
```

Parameters

• **subps** – collection of subparsers in main parser
• **parents** – parsers from which arguments are inherited

The classify subcommand takes specific arguments:

- **--cov_min** (minimum coverage threshold for an edge)
- **--id_min** (minimum identity threshold for an edge)
- **--resolution** (number of identity thresholds to test)

**pyani.scripts.parsers.common_parser module**

Provides parser for arguments common to all subcommands.

```python
def common_parser.build() -> argparse.ArgumentParser:
    Return the common argument parser for all script subcommands.
```

Parameters

• **subps** – collection of subparsers in main parser
• **parents** – parsers from which arguments are inherited

Common arguments are:

- `-l, --logfile` (specify logfile output path) `-v, --verbose` (produce verbose output on STDOUT)

**pyani.scripts.parsers.createdb_parser module**

Provides parser for createdb subcommand.

```python
def createdb_parser.build(subps: argparse._SubParsersAction, parents: Optional[List[argparse.ArgumentParser]] = None) -> None:
    Return a command-line parser for the createdb subcommand.
```

Parameters

• **subps** – collection of subparsers in main parser
• **parents** – parsers from which arguments are inherited
pyani.scripts.parsers.download_parser module

Provides parser for download subcommand.

```python
pyani.scripts.parsers.download_parser.build(subps: argparse._SubParsersAction, parents: Optional[List[argparse.ArgumentParser]] = None) → None
```

Return a command-line parser for the download subcommand.

**Parameters**
- `subps` – ArgumentParser subparser
- `parents` – additional Parser objects

The download subcommand takes specific arguments:

- `-t`, `--taxon` (NCBI taxonomy IDs - comma-separated list, or one ID)
- `--email` (email for providing to Entrez services)
- `--api_key` (path to file containing personal API key for Entrez)
- `--retries` (number of Entrez retry attempts to make)
- `--batchsize` (number of Entrez records to download in a batch)
- `--timeout` (how long to wait for Entrez query timeout)
- `-f`, `--force` (allow existing directory overwrite)
- `--noclobber` (don’t replace existing files)
- `--labels` (path to write labels file)
- `--classes` (path to write classes file)

pyani.scripts.parsers.index_parser module

Provides parser for index subcommand.

```python
pyani.scripts.parsers.index_parser.build(subps: argparse._SubParsersAction, parents: Optional[List[argparse.ArgumentParser]] = None) → None
```

Return a command-line parser for the index subcommand.

**Parameters**
- `subps` – collection of subparsers in main parser
- `parents` – parsers from which arguments are inherited

The index subcommand takes a single positional argument:

- `indir` (directory containing input genome sequence files)

pyani.scripts.parsers.plot_parser module

Provides parser for plot subcommand.

```python
pyani.scripts.parsers.plot_parser.build(subps: argparse._SubParsersAction, parents: Optional[List[argparse.ArgumentParser]] = None) → None
```

Return a command-line parser for the plot subcommand.

**Parameters**
- `subps` – collection of subparsers in main parser
- `parents` – parsers from which arguments are inherited

The plot subcommand takes specific arguments:

- `--method` (graphics method to use)
**pyani.scripts.parsers.report_parser module**

Provides parser for report subcommand.

```python
pyani.scripts.parsers.report_parser.build(subps: argparse._SubParsersAction, parents: Optional[List[argparse.ArgumentParser]] = None) → None
```

Return a command-line parser for the report subcommand.

**Parameters**

- **`subps`** – collection of subparsers in main parser
- **`parents`** – parsers from which arguments are inherited

**pyani.scripts.parsers.run_common_parser module**

Provides parser for arguments common to analysis run subcommands.

```python
pyani.scripts.parsers.run_common_parser.build() → argparse.ArgumentParser
```

Return the common argument parser for analysis run subcommands.

**Parameters**

- **`subps`** – collection of subparsers in main parser
- **`parents`** – parsers from which arguments are inherited

Common arguments are:

- **--name** (human-readable name for the run)
- **--labels** (genome labels for this run)
- **--classes** (genome classes for this run)

**pyani.scripts.parsers.scheduling_parser module**

Provides parser for arguments common to job scheduling.

```python
pyani.scripts.parsers.scheduling_parser.build() → argparse.ArgumentParser
```

Return the common argument parser for job scheduling.

**Parameters**

- **`subps`** – collection of subparsers in main parser
- **`parents`** – parsers from which arguments are inherited

Common arguments are:

**pyani.scripts.subcommands package**

Module providing subcommands for pyani scripts.
Submodules

**pyani.scripts.subcommands.subcmd_anib module**

Provides the anib subcommand for pyani.

```python
```

Return path to fragmented sequence file and JSON of fragment lengths.

**Parameters**

- **inpath** – Path to genome file
- **outdir** – Path to directory to hold fragmented files
- **fragsize** – size of genome fragments

Returns a tuple of (path, json) where path is the path to the fragment file and json is a JSON-ified dictionary of fragment lengths, keyed by fragment sequence ID.

```python
```

Return list of ComparisonJobs.

**Parameters**

- **comparisons** – list of (Genome, Genome) tuples for which comparisons are needed
- **existingfiles** – list of pre-existing BLASTN+ outputs
- **fragfiles** –
- **fraglens** –
- **args** – Namespace, command-line arguments
- **logger** – logging object

```python
```

Perform ANIb on all genome files in an input directory.

**Parameters**

- **args** – Namespace, command-line arguments
- **logger** – logging object


All FASTA format files (selected by suffix) in the input directory are fragmented into (by default 1020nt) consecutive sections, and a BLAST+ database constructed from the whole genome input. The BLAST+ blastn tool is then used to query each set of fragments against each BLAST+ database, in turn.
For each query, the BLAST+ .tab output is parsed to obtain alignment length, identity and similarity error count. Alignments below a threshold are not included in the calculation (this introduces systematic bias with respect to ANIm). The results are processed to calculate the ANI percentages, coverage, and similarity error.

The calculated values are stored in the local SQLite3 database.

**pyani.scripts.subcommands.subcmd_aniblastall module**

Provides the aniblastall subcommand for pyani.

```python
```

Perform ANIblastall on all genome files in an input directory.

**Parameters**

- `args` –
- `logger` –

**pyani.scripts.subcommands.subcmd_anim module**

Provides the anim subcommand for pyani.

```python
class pyani.scripts.subcommands.subcmd_anim.ComparisonJob
    Bases: tuple
    Pairwise comparison job for the SQLAlchemy implementation.
    filtercmd
        Alias for field number 2
    job
        Alias for field number 5
    nucmercmd
        Alias for field number 3
    outfile
        Alias for field number 4
    query
        Alias for field number 0
    subject
        Alias for field number 1

class pyani.scripts.subcommands.subcmd_anim.ComparisonResult
    Bases: tuple
    Convenience struct for a single nucmer comparison result.
    aln_length
        Alias for field number 2
    pid
        Alias for field number 4
```
qcov
    Alias for field number 7
qid
    Alias for field number 0
qlen
    Alias for field number 5
scov
    Alias for field number 8
sid
    Alias for field number 1
sim_errs
    Alias for field number 3
slen
    Alias for field number 6

class pyani.scripts.subcommands.subcmd_anim.ProgData
    Bases: tuple
    Convenience struct for comparison program data/info.
    
    program
        Alias for field number 0
    
    version
        Alias for field number 1

class pyani.scripts.subcommands.subcmd_anim.ProgParams
    Bases: tuple
    Convenience struct for comparison parameters.
    Use default of zero for fragsize or else db queries will not work as SQLite/Python nulls do not match up well
    
    fragsize
        Alias for field number 0
    
    maxmatch
        Alias for field number 1

class pyani.scripts.subcommands.subcmd_anim.RunData
    Bases: tuple
    Convenience struct describing an analysis run.
    
    cmdline
        Alias for field number 3
    
    date
        Alias for field number 2
    
    method
        Alias for field number 0
    
    name
        Alias for field number 1

Chapter 5. Reporting problems and requesting improvements
pyani.scripts.subcommands.subcmd_anim.generate_joblist (comparisons:
    List[Tuple], existing-
    files: List[pathlib.Path],
    args: argparse.Namespace,
    logger: logging.Logger) →
    List[pyani.scripts.subcommands.subcmd_anim.ComparisonJob]

Return list of ComparisonJobs.

Parameters

- **comparisons** – list of (Genome, Genome) tuples
- **existingfiles** – list of pre-existing nucmer output files
- **args** – Namespace of command-line arguments for the run
- **logger** – logging object

pyani.scripts.subcommands.subcmd_anim.run_anim_jobs (joblist:
    List[pyani.scripts.subcommands.subcmd_anim.ComparisonJob],
    args: argparse.Namespace, log-
    ger: logging.Logger) → None

Pass ANIm nucmer jobs to the scheduler.

Parameters

- **joblist** – list of ComparisonJob namedtuples
- **args** – command-line arguments for the run
- **logger** – logging output

pyani.scripts.subcommands.subcmd_anim.subcmd_anim (args: argparse.Namespace,
    logger: logging.Logger) → None

Perform ANIm on all genome files in an input directory.

Parameters

- **args** – Namespace, command-line arguments
- **logger** – logging object

Finds ANI by the ANIm method, as described in Richter et al (2009) Proc Natl Acad Sci USA 106: 19126-19131

All FASTA format files (selected by suffix) in the input directory are compared against each other, pairwise,
using NUCmer (whose path must be provided).

For each pairwise comparison, the NUCmer .delta file output is parsed to obtain an alignment length and simi-
larity error count for every unique region alignment between the two organisms, as represented by sequences in
the FASTA files. These are processed to calculated aligned sequence lengths, average nucleotide identity (ANI)
percentages, coverage (aligned percentage of whole genome - forward direction), and similarity error count for
each pairwise comparison.

The calculated values are deposited in the SQLite3 database being used for the analysis.

For each pairwise comparison the NUCmer output is stored in the output directory for long enough to extract
summary information, but for each run the output is gzip compressed. Once all runs are complete, the outputs
for each comparison are concatenated into a single gzip archive.

5.12. pyani package 57
Update the Comparison table with the completed result set.

**Parameters**

- **joblist** – list of ComparisonJob namedtuples
- **run** – Run ORM object for the current ANIm run
- **session** – active pyanidb session via ORM
- **nucmer_version** – version of nucmer used for the comparison
- **args** – command-line arguments for this run
- **logger** – logging output

The Comparison table stores individual comparison results, one per row.

**pyani.scripts.subcommands.subcmd_classify module**

Provides the classify subcommand for pyani.

```python
class pyani.scripts.subcommands.subcmd_classify.SubgraphData
    Bases: tuple
    Subgraph clique/classification output.
    cliqueinfo
        Alias for field number 2
    graph
        Alias for field number 1
    interval
        Alias for field number 0
```

Generate classifications for an analysis.

**Parameters**

- **args** – Namespace, command-line arguments
- **logger** – logging object
Return graphs trimmed from lowest to highest attribute value.

**Parameters**

- **ingraph** – nx.Graph of genomes as nodes, having edges weighted by the property named in attribute
- **args** – Namespace, parsed command-line arguments
- **attribute** – str, name of the property by which the graph edges should be trimmed

A generator which, starting from the initial graph, yields in sequence a series of graphs from which the edge(s) with the lowest threshold value attribute were removed. The generator returns a tuple of:

(threshold, graph, analyse_cliques(graph))

This will be slow with moderate-large graphs

### pyani.scripts.subcommands.subcmd_createdb module

Provides the createdb subcommand for pyani.

**pyani.scripts.subcommands.subcmd_createdb.subcmd_createdb**

Create an empty pyani database.

**Parameters**

- **args** – Namespace, command-line arguments
- **logger** – logging object

### pyani.scripts.subcommands.subcmd_download module

Provides the download subcommand for pyani.

**pyani.scripts.subcommands.subcmd_download.subcmd_download**

Download assembled genomes in subtree of passed NCBI taxon ID.

**Parameters**

- **args** – Namespace, command-line arguments
- **logger** – logging object
pyaniDocumentation, Release 0.2.9

pyani.scripts.subcommands.subcmd_index module

Provides the index subcommand for pyani.

pyani.scripts.subcommands.subcmd_index.subcmd_index (args: argparse.Namespace, logger: logging.Logger) → int

Generate a file with the MD5 hash for each genome in an input directory.

Parameters

- **args** – Namespace, received command-line arguments
- **logger** – logging object

Identify the genome files in the input directory, and generate a single MD5 for each so that <genome>.fna produces <genome>.md5

Genome files (FASTA) are identified from the file extension.

pyani.scripts.subcommands.subcmd_plot module

Provides the plot subcommand for pyani.

pyani.scripts.subcommands.subcmd_plot.subcmd_plot (args: argparse.Namespace, logger: logging.Logger) → int

Produce graphical output for an analysis.

Parameters

- **args** – Namespace of command-line arguments
- **logger** – logging object

This is graphical output for representing the ANI analysis results, and takes the form of a heatmap, or heatmap with dendrogram.


Write distribution plots for each matrix type.

Parameters

- **run_id** – int, run_id for this run
- **matdata** – MatrixData object for this distribution plot
- **args** – Namespace for command-line arguments
- **outfmts** – list of output formats for files
- **logger** – logging object

Write a single heatmap for a pyani run.

**Parameters**

- **run_id** – int, run_id for this run
- **matdata** – MatrixData object for this heatmap
- **result_labels** – dict of result labels
- **result_classes** – dict of result classes
- **args** – Namespace for command-line arguments
- **outfmts** – list of output formats for files
- **logger** – logging object

```python
```

Write all heatmaps for a specified run to file.

**Parameters**

- **run_id** – int, run identifier in database session
- **session** – Session, active SQLite session
- **outfmts** – list of output format types
- **args** – Namespace, command line arguments
- **logger** – logging object

```python
```

**pyani.scripts.subcommands.subcmd_report module**

Provides the report subcommand for pyani.

```python
class pyani.scripts.subcommands.subcmd_report.ReportParams
    Bases: tuple
    Report query/header data.
    headers
        Alias for field number 2
    name
        Alias for field number 0
    statement
        Alias for field number 1

```

Return processed list of output formats for writing reports.

**Parameters** **args** – Namespace of command-line arguments
Write tabular report of pyani runs from database.

Parameters

- **args** – Namespace of command-line arguments
- **logger** – logging object
- **session** – SQLAlchemy database session
- **formats** – list of output formats
- **params** – ReportParams namedtuple

Present report on ANI results and/or database contents.

Parameters

- **args** – Namespace, command-line arguments
- **logger** – logging object

The report subcommand takes any of several long options that do one of two things:

1. perform a single action.
2. set a parameter/format

These will typically take an output path to a file or directory into which the report will be written (whatever form it takes). By default, text output is written in plain text format, but for some outputs this can be modified by an ‘excel’ or ‘html’ format specifier, which writes outputs in that format, where possible.

Submodules

**pyani.scripts.average_nucleotide_identity module**

Script that calculates ANI measures for a directory of genomes.

This script calculates Average Nucleotide Identity (ANI) according to one of a number of alternative methods described in, e.g.


ANI is proposed to be the appropriate in silico substitute for DNA-DNA hybridisation (DDH), and so useful for delineating species boundaries. A typical percentage threshold for species boundary in the literature is 95% ANI (e.g. Richter et al. 2009).

All ANI methods follow the basic algorithm:

- Align the genome of organism 1 against that of organism 2, and identify the matching regions
• Calculate the percentage nucleotide identity of the matching regions, as an average for all matching regions

Methods differ on: (1) what alignment algorithm is used, and the choice of parameters (this affects the aligned region boundaries); (2) what the input is for alignment (typically either fragments of fixed size, or the most complete assembly available); (3) whether a reciprocal comparison is necessary or desirable.

ANIm: uses MUMmer (NUCmer) to align the input sequences. ANIb: uses BLASTN to align 1000nt fragments of the input sequences TETRA: calculates tetrancleotide frequencies of each input sequence

This script takes as main input a directory containing a set of correctly-formatted FASTA multiple sequence files. All sequences for a single organism should be contained in only one sequence file. The names of these files are used for identification, so it would be advisable to name them sensibly.

Output is written to a named directory. The output files differ depending on the chosen ANI method.

ANIm: MUMmer/NUCmer .delta files, describing the sequence alignment; tab-separated format plain text tables describing total alignment lengths, and total alignment percentile identity

ANIb: FASTA sequences describing 1000nt fragments of each input sequence; BLAST nucleotide databases - one for each set of fragments; and BLASTN output files (tab-separated tabular format plain text) - one for each pairwise comparison of input sequences. There are potentially a lot of intermediate files.

TETRA: Tab-separated text file describing the Z-scores for each tetrancleotide in each input sequence.

In addition, all methods produce a table of output percentage identity (ANIm and ANIb) or correlation (TETRA), between each sequence.

If graphical output is chosen, the output directory will also contain PDF files representing the similarity between sequences as a heatmap with row and column dendrograms.

DEPENDENCIES

- Biopython (http://www.biopython.org)
- BLAST+ executable in the $PATH, or available on the command line (ANIb) (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/)
- MUMmer executables in the $PATH, or available on the command line (ANIm) (http://mummer.sourceforge.net/)

For graphical output

- R with shared libraries installed on the system, for graphical output (http://cran.r-project.org/)
- Rpy2 (http://rpy.sourceforge.net/rpy2.html)

pyani.scripts.average_nucleotide_identity.build_logger(args: argparse.Namespace, logger: Optional[logging.Logger]) → logging.Logger

Return a logging object for the script.

Parameters

- **args** – Namespace of command-line arguments
- **logger** – Expected to be None, but may be logging object
Return ANIm result dataframes for files in input directory.

Parameters

- **args** – Namespace, command-line arguments
- **logger** – logging object
- **infiles** – list of paths to each input file
- **org_lengths** – dict, input sequence lengths, keyed by sequence


All FASTA format files (selected by suffix) in the input directory are compared against each other, pairwise, using NUCmer (which must be in the path). NUCmer output is stored in the output directory.

The NUCmer .delta file output is parsed to obtain an alignment length and similarity error count for every unique region alignment between the two organisms, as represented by the sequences in the FASTA files.

These are processed to give matrices of aligned sequence lengths, average nucleotide identity (ANI) percentages, coverage (aligned percentage of whole genome), and similarity error count for each pairwise comparison.

Calculate TETRA for files in input directory.

Parameters

- **logger** – logging object
- **infiles** – list, paths to each input file

Calculates TETRA correlation scores, as described in:


and


Compress the contents of the passed directory to .tar.gz and delete.

Draw ANIb/ANIm/TETRA results.

Parameters

- **args** – Namespace, command-line arguments
- **logger** – logging object
- **filestems** –
  - filestems for output files
- **gformat** –
  - the format for output graphics


Return function and config for the chosen method.

Parameters

- **args** – Namespace of command-line arguments
- **logger** – logging object

The dictionary defines pairs of method function and configurations, keyed by method name.

pyani.scripts.average_nucleotide_identity.last_exception() → str

Return last exception as a string, or use in logging.

pyani.scripts.average_nucleotide_identity.make_outdirs(args: argparse.Namespace, logger: logging.Logger)

Make the output directory, if required.

Parameters

- **args** – Namespace of command-line options
- **logger** – logging object

If the output directory already exists and args.force is not set True, stop with an error.

If **args.force is set**... If args.noclobber is not set True, delete the output directory tree If args.noclobber is set True, use the existing output directory, and keep any existing output


Return tuple of fragment files, and fragment sizes.
Parameters

- `args` – Namespace of command-line arguments
- `logger` – logging object
- `infiles` – iterable of sequence files to fragment
- `blastdir` – path of directory to place BLASTN databases of fragments

Splits input FASTA sequence files into the fragments (a requirement for ANIb methods), and writes BLAST databases of these fragments, and fragment lengths of sequences, to local files.

```python
pyani.scripts.average_nucleotide_identity.parse_cmdline(argv: Optional[List[T]] = None) -> argparse.Namespace
```

Parse command-line arguments for script.

**Parameters**

- `argv` – list of arguments from command-line

```python
pyani.scripts.average_nucleotide_identity.process_arguments(args: Optional[argparse.Namespace]) -> argparse.Namespace
```

Process command-line arguments.

**Parameters**

- `args` – Namespace of command-line arguments

Either returns parsed arguments or - if only the script name is used, shows the version and exits.

```python
```

Run BLAST commands for ANIb methods.

**Parameters**

- `args` – Namespace of command-line options
- `logger` – logging object
- `infiles` – iterable of sequence files to compare
- `blastdir` – path of directory to fragment BLASTN databases

Runs BLAST database creation and comparisons, returning the cumulative return values of the BLAST tool subprocesses, and the fragment sizes for each input file.

```python
```

Run main process for average_nucleotide_identity.py script.

**Parameters**

- `argsin` – Namespace, command-line arguments
- `logger` – logging object
Return a random subsample of the passed input files.

Parameters

- **args** – Namespace, command-line arguments
- **logger** – logging object
- **infiles** – list of input files for analysis

Raise error and exit if label and class files exist.

Parameters

- **args** – Namespace of command-line arguments
- **logger** – logging object

Exits if class and label files are not found

Test if the specified scheduler can be used.

Parameters

- **args** – Namespace of command-line arguments
- **logger** – logging object

Exits if the scheduler is invalid

Calculate ANIb for files in input directory.

Parameters

- **args** – Namespace of command-line options
- **logger** – logging object
- **infiles** – iterable of paths to each input file
- **org_lengths** – dict of input sequence lengths keyed by sequence name
Calculates ANI by the ANIb method, as described in Goris et al. (2007) Int J Syst Evol Micr 57: 81-91. doi:10.1099/ijs.0.64483-0. There are some minor differences depending on whether BLAST+ or legacy BLAST (BLASTALL) methods are used.

All FASTA format files (selected by suffix) in the input directory are used to construct BLAST databases, placed in the output directory. Each file’s contents are also split into sequence fragments of length options.fragsize, and the multiple FASTA file that results written to the output directory. These are BLASTNed, pairwise, against the databases.

The BLAST output is interrogated for all fragment matches that cover at least 70% of the query sequence, with at least 30% nucleotide identity over the full length of the query sequence. This is an odd choice and doesn’t correspond to the twilight zone limit as implied by Goris et al. We persist with their definition, however. Only these qualifying matches contribute to the total aligned length, and total aligned sequence identity used to calculate ANI.

The results are processed to give matrices of aligned sequence length (aln_lengths.tab), similarity error counts (sim_errors.tab), ANIs (perc_ids.tab), and minimum aligned percentage (perc_aln.tab) of each genome, for each pairwise comparison. These are written to the output directory in plain text tab-separated format.

```
```

Write ANIb/ANIm/TETRA results to output directory.

**Parameters**

- `args` – Namespace, command-line arguments
- `logger` – logging object
- `results` – Results object from analysis

Each dataframe is written to an Excel-format file (if `args.write_excel` is True), and plain text tab-separated file in the output directory. The order of result output must be reflected in the order of filestems.

```
pyani.scripts.delta_filter_wrapper module
```

Wrapper for MUMmer 3.23 delta-filter script.

It is required in order to catch STDOUT ahead of the SGE job runner so that pyani can run on SGE/OGE scheduling systems.

The wrapper does not modify the output of delta-filter, and is called in exactly the same way, passing arguments through directly. The only departure from this is that the first argument is the path to delta-filter, and the final argument denotes the output filtered delta file path, so that redirection (which no longer works with SGE/OGE) is not necessary.

For example, the delta-filter command
delta-filter [options] <delta file> > <filtered delta file>
becomes
delta_filter_wrapper.py delta-filter [options] <delta file> <filtered delta file>

This wrapper is not very robust, but will be improved in later versions of pyani.

```
pyani.scripts.delta_filter_wrapper.run_main() → int
```

Run main process for delta_filter_wrapper.py.
pyani.scripts.genbank_get_genomes_by_taxon module

Script to download from NCBI all genomes in a specified taxon subtree.
This script takes an NCBI taxonomy identifier (or string, though this is not always reliable for taxonomy tree subgraphs...) and downloads all genomes it can find from NCBI in the corresponding taxon subgraph that has the passed argument as root.

exception pyani.scripts.genbank_get_genomes_by_taxon.NCBIException

Bases: Exception

General exception for failed NCBI download.

pyani.scripts.genbank_get_genomes_by_taxon.entrez_batch_webhistory (args, logger, record, expected, batchsize, *fnargs, **fnkwargs)

Recover Entrez data from a prior NCBI webhistory search.

Parameters
- **args** – Namespace, command-line arguments
- **logger** – logging object
- **record** – Entrez webhistory record
- **expected** – int, number of expected search returns
- **batchsize** – int, number of search returns to retrieve in each batch
- ***fnargs** – tuple, arguments to Efetch
- ****fnkwargs** – dict, keyword arguments to Efetch

Recovery is performed in in batches of defined size, using Efetch. Returns all results as a list.

pyani.scripts.genbank_get_genomes_by_taxon.entrez_retry (args, logger, func, *fnargs, **fnkwargs)

Retry the passed function a defined number of times.

Parameters
- **args** – Namespace, command-line arguments
- **logger** – logging object
- **func** – func, Entrez function to attempt
- ***fnargs** – tuple, arguments to the Entrez function
- ****fnkwargs** – dict, keyword arguments to the Entrez function

pyani.scripts.genbank_get_genomes_by_taxon.extract_archive (archivepath, logger)

Return path to extracted gzip file.

Parameters
- **archivepath** – Path, path to gzipped file with “.tar.gz” suffix
- **logger** – logging object
pyani.scripts.genbank_get_genomes_by_taxon.extract_filestem(data)
Extract filestem from Entrez eSummary data.

    Parameters  data – Entrez eSummary

    Function expects esummary[‘DocumentSummarySet’][‘DocumentSummary’][0]
Some illegal characters may occur in AssemblyName - for these, a more robust regex replace/escape may be
required. Sadly, NCBI don’t just use standard percent escapes, but instead replace certain characters with
underscores: white space, slash, comma, hash, brackets.

pyani.scripts.genbank_get_genomes_by_taxon.get_asm_uids(args, logger, taxon_uid)
Return a set of NCBI UIDs associated with the passed taxon.

    Parameters

    • args – Namespace, command-line arguments
    • logger – logging object
    • taxon_uid – str, NCBI taxon ID

    This query at NCBI returns all assemblies for the taxon subtree rooted at the passed taxon_uid.

pyani.scripts.genbank_get_genomes_by_taxon.get_ncbi_asm(args, logger, asm_uid, fmt='fasta')
Return the NCBI AssemblyAccession and AssemblyName for an assembly.

    Parameters

    • args – Namespace, command-line arguments
    • logger – logging object
    • asm_uid – NCBI assembly UID
    • fmt – str, format to retrieve assembly information

    Returns organism data for class,label files also, as well as accession, so we can track whether downloads fail
because only the most recent version is available. .

AssemblyAccession and AssemblyName are data fields in the eSummary record, and correspond to downloadable
where <AA> is AssemblyAccession, and <AN> is AssemblyName, and the choice of GCA vs GCF, and the
three values of nnn are taken from <AA>

pyani.scripts.genbank_get_genomes_by_taxon.last_exception()
Return last exception as a string, or use in logging.

pyani.scripts.genbank_get_genomes_by_taxon.logreport_downloaded(accn, skiplist, accndict, uidaccndict, logger)
Report to logger if alternative assemblies were downloaded.

    Parameters

    • accn –
    • skiplist –
    • accndict –
    • uidaccndict –
    • logger – logging object
pyani.scripts.genbank_get_genomes_by_taxon.make_outdir(args, logger)

Make the output directory, if required.

Parameters

- **args** – Namespace, command-line arguments
- **logger** – logging object

This is a little involved. If the output directory already exists, we take the safe option by default, and stop with an error. We can, however, choose to force the program to go on, in which case we can either clobber the existing directory, or not. The options turn out as the following, if the directory exists: DEFAULT: stop and report the collision FORCE: continue, and remove the existing output directory NOCLOBBER+FORCE: continue, but do not remove the existing output

pyani.scripts.genbank_get_genomes_by_taxon.parse_cmdline(argv=None)

Parse command-line arguments.

Parameters **argv** – list of command-line arguments


Download assembly sequence to a local directory.

Parameters

- **args** – Namespace, command-line arguments
- **logger** – logging object
- **filestem** – str, filestem for output file
- **ftppstem** – str, URI stem for NCBI FTP site
- **fmt** – str, format for output file

The filestem corresponds to <AA>_<AN>, where <AA> and <AN> are AssemblyAccession and AssemblyName: data fields in the eSummary record. These correspond to downloadable files for each assembly at ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GC{AF}/nnn/nnn/nnn/<AA>_<AN>/ where <AA> is AssemblyAccession, and <AN> is AssemblyName. The choice of GCA vs GCF, and the values of nnn, are derived from <AA>

The files in this directory all have the stem <AA>_<AN>_<suffix>, where suffixes are: assembly_report.txt assembly_stats.txt feature_table.txt.gz genomic.fna.gz genomic.gbff.gz genomic.gff.gz protein.faa.gz protein.gpff.gz rm_out.gz rm.run wgsmaster.gbff.gz

This function downloads the genomic_fna.gz file, and extracts it in the output directory name specified when the script is called.

pyani.scripts.genbank_get_genomes_by_taxon.run_main(args=None, logger=None)

Run main process for average_nucleotide_identity.py script.

Parameters

- **args** – Namespace, command-line arguments
- **logger** – logging object

pyani.scripts.genbank_get_genomes_by_taxon.set_ncbi_email(args, logger)

Set contact email for NCBI.

Parameters

- **args** – Namespace, command-line arguments
pyani.scripts.genbank_get_genomes_by_taxon.write_contigs(args, logger, asm_uid, contig_uids, batch_size=10000)

Write assembly contigs to a single FASTA file.

Parameters

- **args** – Namespace, command-line arguments
- **logger** – logging object
- **asm_uid** – str, NCBI assembly UID
- **contig_uids** –
- **batch_size** – int

FASTA records are returned, as GenBank and even GenBankWithParts format records don’t reliably give correct sequence in all cases.

The script returns two strings for each assembly, a ‘class’ and a ‘label’ string - this is for use with, e.g. pyani.

---

**pyani.scripts.logger module**

Module providing a script logger.

This wraps the builtin logging module classes

```python
pyani.scripts.logger.build_logger(name: str, args: argparse.Namespace) → logging.Logger
```

Return a logger for this script.

Parameters

- **name** – str, name for logger
- **args** – Namespace, command-line arguments

Instantiates a logger for the script, and adds basic info.

---

**pyani.scripts.pyani_script module**

Implements the pyani script for classifying prokaryotic genomes.

```python
pyani.scripts.pyani_script.run_main(argv: Optional[List[str]] = None, logger: Optional[logging.Logger] = None) → int
```

Run main process for pyani.py script.

Parameters

- **argv** –
- **logger** –

---

**pyani.scripts.tools module**

Module providing functions in support of the pyani command-line scripts.
class pyani.scripts.tools.DLFileData
    Bases: tuple
    Convenience struct for file download data.

    filstem
        Alias for field number 0

    ftpstem
        Alias for field number 1

    suffix
        Alias for field number 2

exception pyani.scripts.tools.PyaniScriptException(msg='Error in pyani.py script')
    Bases: Exception
    General exception for pyani.py script.

    Download genome and accompanying MD5 hash from NCBI.

    Parameters
    • args – Namespace for command-line arguments
    • logger – logging object
    • dlfiledata – namedtuple of info for file to download
    • dltlate – reference database to use: RefSeq or GenBank
    • disable_tqdm – disable progress bar

    This function tries the (assumed to be passed) RefSeq FTP URL first and, if that fails, then attempts to download
    the corresponding GenBank data.

    We attempt to gracefully skip genomes with download errors.

pyani.scripts.tools.make_asm_dict(taxon_ids: List[str], retries: int) → Dict[KT, VT]
    Return a dict of assembly UIDs, keyed by each passed taxon ID.

    Parameters
    • taxon_ids –
    • retries –

    Create output directory (allows for force and noclobber).

    Parameters
    • outdir – Path, path to output directory
    • force –
    • noclobber –
    • logger –
The intended outcomes are: outdir doesn’t exist: create outdir outdir exists: raise exception outdir exists, –force only: remove the directory tree outdir exists, –force –noclobber: continue with existing directory tree but do not overwrite files

So long as the outdir is created with this function, we need only check for args.noclobber elsewhere to see how to proceed when a file exists.

5.12.2 Submodules

pyani.anib module

Code to implement the ANIb average nucleotide identity method.

Calculates ANI by the ANIb method, as described in Goris et al. (2007) Int J Syst Evol Micro 57: 81-91. doi:10.1099/ijs.0.64483-0.

From Goris et al.

‘‘The genomic sequence from one of the genomes in a pair (the query) was cut into consecutive 1020 nt fragments. The 1020 nt cut-off was used to correspond with the fragmentation of the genomic DNA to approximately 1 kb fragments during the DDH experiments. […] The 1020 nt fragments were then used to search against the whole genomic sequence of the other genome in the pair (the reference) by using the BLASTN algorithm; the best BLASTN match was saved for further analysis. The BLAST algorithm was run using the following settings: X=150 (where X is the drop-off value for gapped alignment), q=-1 (where q is the penalty for nucleotide mismatch) and F=F (where F is the filter for repeated sequences); the rest of the parameters were used at the default settings. These settings give better sensitivity than the default settings when more distantly related genomes are being compared, as the latter target sequences that are more similar to each other. […] The ANI between the query genome and the reference genome was calculated as the mean identity of all BLASTN matches that showed more than 30% overall sequence identity (recalculated to an identity along the entire sequence) over an alignable region of at least 70% of their length. This cut-off is above the ‘twilight zone’ of similarity searches in which an inference of homology is error prone because of low levels of Reverse searching, i.e. in which the reference genome is used as the query, was also performed to provide reciprocal values.’’

All input FASTA format files are used to construct BLAST databases. Each file’s contents are also split into sequence fragments of length options.fragsize, and the multiple FASTA file that results written to the output directory. These are BLASTNed, pairwise, against the databases.

BLAST output is interrogated for all fragment matches that cover at least 70% of the query sequence, with at least 30% nucleotide identity over the full length of the query sequence. This is an odd choice and doesn’t correspond to the twilight zone limit as implied by Goris et al. We persist with their definition, however. Only these qualifying matches contribute to the total aligned length, and total aligned sequence identity used to calculate ANI.

pyani.anib.build_db_jobs (infiles: List[pathlib.Path], blastcmds: pyani.pyani_tools.BLASTcmds) → Dict[KT, VT]

Return dictionary of db-building commands, keyed by dbname.

Parameters

• infiles

• blastcmds


Return single blastall command.

Parameters

• fname1

Chapter 5. Reporting problems and requesting improvements
• fname2
• outdir
• blastall_exe – str, path to BLASTALL executable


Return a single blastn command.

Parameters
• fname1
• fname2
• outdir
• blastn_exe – str, path to blastn executable


Return formatdb command and path to output file.

Parameters
• filename – Path, input filename
• outdir – Path, path to output directory
• blastdb_exe – Path, path to the formatdb executable


Return makeblastdb command and path to output file.

Parameters
• filename – Path, input filename
• outdir – Path, directory for output
• blastdb_exe – Path, path to the makeblastdb executable


Chop sequences of the passed files into fragments, return filenames.

Parameters
• infiles – collection of paths to each input sequence file
• outdirname – Path, path to output directory
• fragsize – Int, the size of sequence fragments

Takes every sequence from every file in infiles, and splits them into consecutive fragments of length fragsize, (with any trailing sequences being included, even if shorter than fragsize), writing the resulting set of sequences to a file with the same name in the specified output directory.

All fragments are named consecutively and uniquely (within a file) as fragNNNNN. Sequence description fields are retained.
Returns a tuple `(filenames, fragment_lengths)` where `filenames` is a list of paths to the fragment sequence files, and `fragment_lengths` is a dictionary of sequence fragment lengths, keyed by the sequence files, with values being a dictionary of fragment lengths, keyed by fragment IDs.

```python
pyani.anib.generate_blastdb_commands(filenames: List[pathlib.Path], outdir: pathlib.Path,
```

Return list of makeblastdb command-lines for ANIb/ANIblastall.

**Parameters**

- `filenames` – a list of paths to input FASTA files
- `outdir` – path to output directory
- `blastdb_exe` – path to the makeblastdb executable
- `mode` – str, ANIb analysis type (ANIb or ANIblastall)

```python
pyani.anib.generate_blastn_commands(filenames: List[pathlib.Path], outdir: pathlib.Path,
blast_exe: Optional[pathlib.Path] = None, mode: str = 'ANIb') → List[str]
```

Return a list of blastn command-lines for ANIm.

**Parameters**

- `filenames` – a list of paths to fragmented input FASTA files
- `outdir` – path to output directory
- `blastn_exe` – path to BLASTN executable
- `mode` – str, analysis type (ANIb or ANIblastall)

Assumes that the fragment sequence input filenames have the form ACCESSION-fragments.ext, where the corresponding BLAST database filenames have the form ACCESSION.ext. This is the convention followed by the fragment_FASTA_files() function above.

```python
pyani.anib.get_fraglength_dict(fastafiles: List[pathlib.Path]) → Dict[KT, VT]
```

Return dictionary of sequence fragment lengths, keyed by query name.

**Parameters**

- `fastafiles` – list of paths to FASTA input whole sequence files

Loops over input files and, for each, produces a dictionary with fragment lengths, keyed by sequence ID. These are returned as a dictionary with the keys being query IDs derived from filenames.

```python
pyani.anib.get_fragment_lengths(fastafile: pathlib.Path) → Dict[KT, VT]
```

Return dictionary of sequence fragment lengths, keyed by fragment ID.

**Parameters**

- `fastafile` –

Biopython’s SeqIO module is used to parse all sequences in the FASTA file.

NOTE: ambiguity symbols are not discounted.

```python
pyani.anib.get_version(blast_exe: pathlib.Path = PosixPath('blastn')) → str
```

Return BLAST+ blastn version as a string.

**Parameters**

- `blast_exe` – path to blastn executable

We expect blastn to return a string as, for example

```
$ blastn -version
blastn: 2.9.0+
Package: blast 2.9.0, build Jun 10 2019 09:40:53
```
This is concatenated with the OS name.

```python
```

Return BLASTcmds object for construction of BLAST commands.

**Parameters**

- `mode` – str, the kind of ANIb analysis (ANIb or ANIblastall)
- `outdir` –
- `format_exe` –
- `blast_exe` –
- `prefix` –

```python
```

Return job dependency graph, based on the passed input sequence files.

**Parameters**

- `infiles` – list of paths to input FASTA files
- `fragfiles` – list of paths to fragmented input FASTA files
- `blastcmds` –

By default, will run ANIb - it is possible to make a mess of passing the wrong executable for the mode you’re using.

All items in the returned graph list are BLAST executable jobs that must be run after the corresponding database creation. The Job objects corresponding to the database creation are contained as dependencies. How those jobs are scheduled depends on the scheduler (see run_multiprocessing.py, run_sge.py)

```python
pyani.anib.parse_blast_tab(filename: pathlib.Path, fraglengths: Dict[KT, VT], mode: str = 'ANIb') → Tuple[int, int, int]
```

Return (alignment length, similarity errors, mean_pid) tuple.

**Parameters**

- `filename` – Path, path to .blast_tab file
- `fraglengths` – Optional[Dict], dictionary of fragment lengths for each genome.
- `mode` – str, analysis type (ANIb or ANIblastall)

Calculate the alignment length and total number of similarity errors (as we would with ANIm), as well as the Goris et al.-defined mean identity of all valid BLAST matches for the passed BLASTALL alignment .blast_tab file.

‘‘ANI between the query genome and the reference genome was calculated as the mean identity of all BLASTN matches that showed more than 30% overall sequence identity (recalculated to an identity along the entire sequence) over an alignable region of at least 70% of their length. ‘‘

```python
```

Return tuple of ANIb results for .blast_tab files in the output dir.

**Parameters**

- `blast_dir` – Path, path to the directory containing .blast_tab files
• **org_lengths** – Dict, the base count for each input sequence

• **fraglengths** – dictionary of query sequence fragment lengths, only needed for BLASTALL output

• **mode** – str, analysis type (ANIb or ANIblastall)

• **logger** – a logger for messages

Returns the following pandas dataframes in an ANIResults object; query sequences are rows, subject sequences are columns:

- **alignment_lengths** - non-symmetrical: total length of alignment
- **percentage_identity** - non-symmetrical: ANIb (Goris) percentage identity
- **alignment_coverage** - non-symmetrical: coverage of query
- **similarity_errors** - non-symmetrical: count of similarity errors

May throw a ZeroDivisionError if one or more BLAST runs failed, or a very distant sequence was included in the analysis.

**pyani.anim module**

Code to implement the ANIm average nucleotide identity method.


All input FASTA format files are compared against each other, pairwise, using NUCmer (binary location must be provided). NUCmer output will be stored in a specified output directory.

The NUCmer .delta file output is parsed to obtain an alignment length and similarity error count for every unique region alignment. These are processed to give matrices of aligned sequence lengths, similarity error counts, average nucleotide identity (ANI) percentages, and minimum aligned percentage (of whole genome) for each pairwise comparison.

```python
```

Return a tuple of corresponding NUCmer and delta-filter commands.

**Parameters**

- **fname1** – path to query FASTA file
- **fname2** – path to subject FASTA file
- **outdir** – path to output directory
- **nucmer_exe** –
- **filter_exe** –
- **maxmatch** – Boolean flag indicating whether to use NUCmer’s -maxmatch option. If not, the -mum option is used instead

The split into a tuple was made necessary by changes to SGE/OGE. The delta-filter command must now be run as a dependency of the NUCmer command, and be wrapped in a Python script to capture STDOUT.

**NOTE**: This command-line writes output data to a subdirectory of the passed outdir, called “nucmer_output”.

---

Chapter 5. Reporting problems and requesting improvements

Return list of NUCmer command-lines for ANIm.

Parameters

- **filenames** – a list of paths to input FASTA files
- **outdir** – path to output directory
- **nucmer_exe** – location of the nucmer binary
- **maxmatch** – Boolean flag indicating to use NUCmer’s -maxmatch option

The first element returned is a list of NUCmer commands, and the second a corresponding list of delta_filter_wrapper.py commands. The NUCmer commands should each be run before the corresponding delta-filter command.

TODO: This return value needs to be reworked as a collection.

Loop over all FASTA files generating NUCmer command lines for each pairwise comparison.


Return list of Jobs describing NUCmer command-lines for ANIm.

Parameters

- **filenames** – Iterable, Paths to input FASTA files
- **outdir** – str, path to output directory
- **nucmer_exe** – str, location of the nucmer binary
- **filter_exe** –
- **maxmatch** – Boolean flag indicating to use NUCmer’s -maxmatch option
- **jobprefix** –

Loop over all FASTA files, generating Jobs describing NUCmer command lines for each pairwise comparison.

pyani.anim.get_fasta_files(dirname: pathlib.Path = PosixPath('.')) → Iterable[T_co]

Return iterable of FASTA files in the passed directory.

Parameters **dirname** – str, path to input directory

pyani.anim.get_version(nucmer_exe: pathlib.Path = PosixPath('nucmer')) → str

Return NUCmer package version as a string.

Parameters **nucmer_exe** – path to NUCmer executable

We expect NUCmer to return a string on STDERR as

```
$ nucmer
NUCmer (NUCLEotide MUMmer) version 3.1
```

we concatenate this with the OS name.

pyani.anim.parse_delta(filename: pathlib.Path) → Tuple[int, int]

Return (alignment length, similarity errors) tuple from passed .delta.
Parameters **filename** – Path, path to the input .delta file

Extracts the aligned length and number of similarity errors for each aligned uniquely-matched region, and returns the cumulative total for each as a tuple.

Delta file format has seven numbers in the lines of interest: - start on query - end on query - start on target - end on target - error count (non-identical, plus indels) - similarity errors (non-positive match scores) - stop codons (always zero for nucmer)

`pyani.anim.process_deltadir`(delta_dir: pathlib.Path, org_lengths: Dict[KT, VT], logger: Optional[logging.Logger] = None) → pyani.pyani_tools.ANIResults

Return tuple of ANIm results for .deltas in passed directory.

Parameters

- **delta_dir** – Path, path to the directory containing .delta files
- **org_lengths** – dictionary of total sequence lengths, keyed by sequence

Returns the following pandas dataframes in an ANIResults object; query sequences are rows, subject sequences are columns:

- alignment_lengths - symmetrical: total length of alignment
- percentage_identity - symmetrical: percentage identity of alignment
- alignment_coverage - non-symmetrical: coverage of query and subject
- similarity_errors - symmetrical: count of similarity errors

May throw a ZeroDivisionError if one or more NUCmer runs failed, or a very distant sequence was included in the analysis.

**pyani.blast module**

Code for handling BLAST output files.

`pyani.blast.parse_blasttab`(fhandle: TextIO) → List[List[str]]

Return the passed BLAST tab output file as a list of lists.

Parameters **fhandle** – TextIO, filehandle containing BLAST output file

This is used when testing for conserved BLAST output, as the exact format of the BLAST result can depend on the software version. For instance, the locally-installed version may be BLASTN+ 2.6.0, which reports match identity to 3sf, and the version in CI may be BLASTN+ 2.2.28, which reports to 2sf.

Returning a list of lines, parsed into the appropriate data type, allows for direct comparison of line content independent of formatting.

**pyani.download module**

Module providing functions useful for downloading genomes from NCBI.

```python
class pyani.download.ASMIDs
Bases: tuple

Matching Assembly ID information for a query taxID.

asm_ids
    Alias for field number 2

query
    Alias for field number 0
```
result_count
  Alias for field number 1

class pyani.download.Classification
  Bases: tuple
  Taxonomic classification for an isolate.

  genus
    Alias for field number 1

  organism
    Alias for field number 0

  species
    Alias for field number 2

  strain
    Alias for field number 3

class pyani.download.FileExistsException
  Bases: Exception
  msg: str = 'Specified file exists'

exception pyani.download.FileExistsException

exception pyani.download.NCBIDownloadException
  Bases: Exception
  msg: str = 'Error downloading file from NCBI'

class pyani.download.Hashstatus
  Bases: tuple
  Status report on file hash comparison.

  filehash
    Alias for field number 2

  localhash
    Alias for field number 1

  passed
    Alias for field number 0

exception pyani.download.NCBIDownloadException

pyani.download.check_hash
  Check MD5 of passed file against downloaded NCBI hash file.

Parameters

  • fname – Path, path to local hash file

  • hashfile – Path, path to NCBI hash file

pyani.download.compile_url
  (filestem: str, suffix: str, ftpstem: str) → Tuple[str, str]
  Compile download URLs given a passed filestem.

Parameters

  • filestem –

  • suffix –

  • ftpstem –
The filestem corresponds to `<AA>_<AN>`, where `<AA>` and `<AN>` are AssemblyAccession and AssemblyName: data fields in the eSummary record. These correspond to downloadable files for each assembly at ftp://ftp.ncbi.nlm.nih.gov/ genomes/all/GC{[}AF{]}/nnn/nnn/<AA>_<AN>/ where `<AA>` is AssemblyAccession, and `<AN>` is AssemblyName. The choice of GCA vs GCF, and the values of nnn, are derived from `<AA>`.

The files in this directory all have the stem `<AA>_<AN>_<suffix>`, where suffixes are: assembly_report.txt assembly_stats.txt feature_table.txt.gz genomic.fna.gz genomic.gbff.gz genomic.gff.gz protein.faa.gz protein.gpff.gz rm_out.gz rm.run wgsmaster.gbff.gz


Construct paths to output files for genome and hash.

Parameters

- **filestem** – str, output filename stem
- **suffix** – str, output filename suffix
- **outdir** – Path, path to output directory

pyani.download.create_hash (fname: pathlib.Path) → str

Return MD5 hash of the passed file contents.

Parameters **fname** – Path, path to file for hashing

We can ignore the Bandit B303 error as we’re not using the hash for cryptographic purposes.


Return class and label text from UID classification.

Parameters

- **classification** – Classification named tuple (org, genus, species, strain)
- **filestem** – str, filestem of input genome file
- **genomehash** – str, MD5 hash of genome data

The ‘class’ data is the organism as provided in the passed Classification named tuple; the ‘label’ data is genus, species and strain information from the same tuple. The label is intended to be human-readable, the class data to be a genuine class identifier.

Returns a tuple of two strings: (label, class).

The two strings are tab-separated strings: `<HASH>t<FILE>t<CLASS/LABEL>`. The hash is used to help uniquely identify the genome in the database (label/class is unique by a combination of hash and run ID).

pyani.download.download_url (url: str, outfname: pathlib.Path, timeout: int, disable_tqdm: bool = False) → None

Download remote URL to a local directory.

Parameters

- **url** – URL of remote file for download
- **outfname** – Path, path to write output
- **timeout** –
- **disable_tqdm** – Boolean, show tqdm progress bar?

This function downloads the contents of the passed URL to the passed filename, in buffered chunks.
pyani.download.entrez_batch_webhistory(record, expected, batchsize, retries, *fnargs, **fnkwargs) → List[str]

Recover the Entrez data from a prior NCBI webhistory search.

**Parameters**

- `record` – Entrez webhistory record
- `expected` – number of expected search returns
- `batchsize` – how many search returns to retrieve in a batch
- `retries` – int
- `*fnargs` – arguments to Efetch
- `**fnkwargs` – keyword arguments to Efetch

Recover results in batches of defined size, using Efetch. Returns all results as a list.

pyani.download.entrez_retry(func, retries, *fnargs, **fnkwargs)
retry the passed function up to the number of times specified.

**Parameters**

- `func` – function to be executed
- `retries` – int, number of times to retry function execution
- `*fnargs` – optional arguments to passed function
- `**fnkwargs` – optional keyword arguments to passed function


Extract contents of fname to ename using gunzip.

**Parameters**

- `fname` – str, path to input compressed file
- `ename` – str, path to output uncompressed file

Returns status of subprocess call

pyani.download.extract_filestem(esummary) → str

Extract filestem from Entrez eSummary data.

**Parameters**

- `esummary` –

Function expects esummary[‘DocumentSummarySet’][‘DocumentSummary’][0]

Some illegal characters may occur in AssemblyName - for these, a more robust regex replace/escape may be required. Sadly, NCBI don’t just use standard percent escapes, but instead replace certain characters with underscores: white space, slash, comma, hash, brackets.

pyani.download.extract_hash(hashfile: pathlib.Path, name: str) → str

Return MD5 hash from file of name:MD5 hashes.

**Parameters**

- `hashfile` – Path, path to file containing name:MD5 pairs
- `name` – str, name associated with hash

pyani.download.get_asm_uids(taxon_uid: str, retries: int) → pyani.download.ASMIDs

Return set of NCBI UIDs associated with the passed taxon UID.

**Parameters**
• `taxon_uid` – str, NCBI taxID for taxon to download
• `retries` – int, number of download retry attempts

This query at NCBI returns all assemblies for the taxon subtree rooted at the passed taxon_uid.

```python
pyani.download.get_ncbi_classification(esummary) → pyani.download.Classification

Parameters

- `esummary` –
```

Return organism, genus, species, strain info from eSummary data.

```pyani.download.get_ncbi_esummary(asm_uid, retries, api_key=None) → Tuple

Obtain full eSummary info for the passed assembly UID.

Parameters

- `asm_uid` –
- `retries` –
- `api_key` –

```pyani.download.last_exception() → str

Return last exception as a string.

```pyani.download.parse_api_key(api_path: pathlib.Path) → str

Parse the contents of the file for NCBI API key.

Parameters

- `api_path` – Path, location of NCBI API key

This function expects that the indicated file contains a single line (perhaps with a newline) containing no data other than an NCBI API key.


Download genome contigs and MD5 hash data from NCBI.

Parameters

- `filestem` –
- `suffix` –
- `ftpsstem` –
- `outdir` –
- `timeout` –
- `disable_tqdm` – Boolean, show tqdm progress bar?

```pyani.download.set_ncbi_email(email: str) → None

Set contact email for NCBI.

Parameters

- `email` – str, email address to give to Entrez at NCBI

```pyani.download.split_taxa(taxa: str) → List[str]

Return list of taxon ids from the passed comma-separated list.

Parameters

- `taxa` – str, comma-separated list of valid NCBI taxonomy IDs

The function checks the passed taxon argument against a regular expression that permits comma-separated numerical symbols only.
**pyani.nucmer module**

Code for handling NUCmer output files.

```python
class pyani.nucmer.DeltaAlignment(refstart: int, refend: int, qrystart: int, qryend: int, errs: int, simerrs: int, stops: int)
    Bases: object

    Represents a single alignment region and scores for a pairwise comparison.

    Bases: object

    Represents a comparison between two sequences in a .delta file.

    add_alignment(aln: pyani.nucmer.DeltaAlignment) → None
        Add passed alignment to this object.

        Parameters
        aln -- DeltaAlignment object

class pyani.nucmer.DeltaData(name: str, handle: TextIO = None)
    Bases: object

    Class to hold MUMmer/nucmer output “delta” data.

    This is required because the ordering within files differs depending on MUMmer build, for the same version (as evidenced by differences between OSX and Linux builds), and a means of testing for equality of outputs is necessary.

    The output file structure and format is described at http://mummer.sourceforge.net/manual/#nucmeroutput

    Each file is represented as:

    - header: first line of the .delta file, naming the two input comparison files; stored as a tuple (path1, path2), returned as the combined string; the individual files are stored as self._query and self._subject
    - program: name of the MUMmer program that produced the output
    - query: path to the query sequence file
    - subject: path to the subject sequence file

    comparisons
        Comparisons in the .delta file.

    from_delta(handle: TextIO) → None
        Populate the object from the passed .delta or .filter filehandle.

    metadata
        Metadata from the .delta file.

    program
        The MUMmer program used for the comparison.

    query
        Query file for the MUMmer comparison.

    reference
        Reference file for the MUMmer comparison.

    Bases: object

    Represents a single sequence comparison header from a MUMmer .delta file.
```

5.12. pyani package

85
class pyani.nucmer.DeltaIterator (handle: TextIO)
    Bases: object

    Iterator for MUMmer .delta files.
    Returns a stream of DeltaMetadata, DeltaComparison and DeltaAlignment objects when iterated over a filehandle.
    The .delta file structure and format is described at http://mummer.sourceforge.net/manual/#nucmeroutput

class pyani.nucmer.DeltaMetadata
    Bases: object

    Represents the metadata header for a MUMmer .delta file.

pyani.pyani_classify module

Module providing functions to generate clusters/species hypotheses.

class pyani.pyani_classify.Cliquesinfo
    Bases: tuple

    Summary of clique structure.
        all_k_complete
            Alias for field number 2
        n_nodes
            Alias for field number 0
        n_subgraphs
            Alias for field number 1

pyani.pyani_classify.all_components_k_complete (graph: networkx.classes.graph.Graph) → bool
    Return True if all components in passed graph are k-complete.

Parameters
    graph – NetworkX Graph object

pyani.pyani_classify.analyse_cliques (graph: networkx.classes.graph.Graph) → pyani.pyani_classify.Cliquesinfo
    Return Cliquesinfo NamedTuple describing clique data for a graph.

Parameters
    graph – NetworkX Graph object

pyani.pyani_classify.build_graph_from_results (results, label_dict: Dict[int, str], cov_min: float = 0, id_min: float = 0) → networkx.classes.graph.Graph
    Return undirected graph representing the passed ANIResults object.

The passed ANIResults object is converted to an undirected graph where nodes on the graph represent genomes, and edges represent pairwise comparisons having the minimum coverage and identity indicated.

Parameters
    • results –
        – Run object from pyani_orm
    • label_dict – dictionary of genome labels for result matrices the dict is keyed by the index/column values for the results matrices
    • cov_min –
        – minimum coverage for an edge
• **id_min** – minimum identity for an edge

```python
pyani.pyani_classify.k_complete_component_status(graph: networkx.classes.graph.Graph) → List[bool]
```

Return list of Booleans of whether connected components of the graph are k-complete.

**Parameters**
- `graph` – NetworkX Graph object

For each component in the passed graph, a list of Booleans is calculated, representing whether each node has property P: the degree of the node is equal to the number of nodes in that component, minus 1.

The all() gives a Boolean indicating whether all nodes in that component have property P.

```python
pyani.pyani_classify.remove_low_weight_edges(graph: networkx.classes.graph.Graph, threshold: float, attribute: str = 'identity') → Tuple[networkx.classes.graph.Graph, List[T]]
```

Return graph and edgelist where edges having weight < threshold are removed.

**Parameters**
- `graph` – NetworkX Graph
- `threshold` – float, minimum edge weight
- `attribute` – String, attribute to use as weight

### pyani.pyani_config module

Configuration settings for the pyani package.

```python
```

Return colormap parameters for a dataframe.

**Parameters**
- `dataframe` –
- `matname` –

The colormap is dependent on the type of analysis that was done.

```python
pyani.pyani_config.params_mpl(dfm: pandas.core.frame.DataFrame) → Dict[str, Tuple[str, Any, Any]]
```

Return dict of matplotlib parameters, dependent on dataframe.

**Parameters**
- `dfm` –

DEPRECATED FROM v0.3 onwards

### pyani.pyani_files module

Code to handle files for average nucleotide identity calculations.

**Exception**

```python
exception pyani.pyani_files.PyaniFilesException
Bases: pyani.PyaniException
```

Exception raised by pyani when file interaction goes bad.
Return a list of existing output files at dirpath.

**Parameters**

- **dirpath** – Path, path to existing output directory
- **program** – str, name of program to use for comparisons
- **args** – Namespace, command-line arguments for the run

Return a list of (FASTA file, hash file) tuples in passed directory.

**Parameters**

- **dirname** – Path, path to input directory

Raises an IOError if the corresponding hash for a FASTA file does not exist.

Return a list of FASTA files in the passed directory.

**Parameters**

- **dirname** – Path, path to input directory

Return a list of paths to files matching a list of FASTA extensions.

**Parameters**

- **dirname** – Path, path to directory containing input FASTA files
- **extlist** – List, file suffixes for FASTA files

Returns the full path to each file.

Return files in passed directory, filtered by extension.

**Parameters**

- **dirname** – Path, path to input directory
- **ext** – optional iterable of arguments describing permitted file extensions

Return dictionary of sequence lengths, keyed by organism.

**Parameters**

- **fastafilenames** – Iterable[Path], paths to input FASTA files

Biopython’s SeqIO module is used to parse all sequences in the FASTA file corresponding to each organism, and the total base count in each is obtained.

NOTE: ambiguity symbols are not discounted.

Return a dictionary of genome classes or labels keyed by hash.

**Parameters**

- **path** – Path, path to classes or labels file

The expected format of the classes and labels files is: <HASH>t<FILESTEM>t<CLASS>|<LABEL>, where <HASH> is the MD5 hash of the genome data (this is not checked); <FILESTEM> is the path to the genome file (this is intended to be a record for humans to audit, it’s not needed for the database interaction; and <CLASS>|<LABEL> is the class or label associated with that genome.
**pyani.pyani_files**

`read_fasta_description(filename: pathlib.Path) → str`

Return the first description string from a FASTA file.

**Parameters**

- `filename` – Path, path to FASTA file

`read_hash_string(filename: pathlib.Path) → Tuple[str, str]`

Return the hash and file strings from the passed hash file.

**Parameters**

- `filename` – Path, path to file containing hash string

**pyani.pyani_graphics module**

Code to implement graphics output for ANI analyses.

**class pyani.pyani_graphics.Params**(params: Tuple, labels: Optional[Dict[KT, VT]] = None, classes: Optional[Dict[KT, VT]] = None)

Bases: object

Convenience class to hold heatmap rendering parameters.

- `vdiff` – Return difference between max and min values for presentation.

**pyani.pyani_jobs module**

Code to manage jobs for pyani.

In order to be a little more consistent behind the scenes for schedulers, and to allow for a fairly hacky approach to scheduling on SGE, a job dependency graph is used.

Commands to be run are stored in Jobs. A Job’s dependency is stored so that the Job will not be executed until its dependency is executed.

When used in ANI analysis, the way jobs are used depends on the scheduler.

With multiprocessing, we place all root jobs in a single pool; then all first-level dependencies will go in a second (dependent) pool that is not run until the first is completed, and so on. It’s not very efficient, but should work equivalently to the original code that handled asynchronous pools directly.

With SGE, the dependencies can be managed independently, and effectively interleaved by the scheduler with no need for pools.

This code is essentially a frozen and cut-down version of pysge (https://github.com/widdowquinn/pysge)


Bases: object

Individual job to be run, with list of dependencies.

- `add_dependency(job) → None`
  
  Add passed job to the dependency list for this Job.

  **Parameters**

  - `job` – Job to be added to the Job’s dependency list

  This Job should not execute until all dependent jobs are completed.

- `remove_dependency(job) → None`

  Remove passed job from this Job’s dependency list.

  **Parameters**

  - `job` – Job to be removed from the Job’s dependency list

- `wait(interval: float = 0.01) → None`

  Wait until the job finishes, and poll SGE on its status.
**Parameters**

- **interval** – float, number of seconds to wait before polling SGE

**Class** `pyani.pyani_jobs.JobGroup`

```python
```

- **Bases**: `object`

Class that stores a group of jobs, permitting parameter sweeps.

- **add_dependency** *(job) → None*

  Add the passed job to the dependency list for this JobGroup.

  - **Parameters**
    - **job** – Job, job to be added to the JobGroup’s dependency list

- **generate_script** *( ) → None*

  Create the SGE script that will run the jobs in the JobGroup.

- **remove_dependency** *(job) → None*

  Remove passed job from this JobGroup’s dependency list.

  - **Parameters**
    - **job** – Job, job to be removed from the JobGroup’s dependency list

- **wait** *(interval: float = 0.01) → None*

  Wait for a defined period, then poll SGE for job status.

  - **Parameters**
    - **interval** – int, seconds to wait before polling SGE

**pyani.pyani_orm module**

Module providing useful functions for manipulating `pyani`’s SQLite3 db.

This SQLAlchemy-based ORM replaces the previous SQL-based module

```python
class pyani.pyani_orm.BlastDB(**kwargs)
```

- **Bases**: `sqlalchemy.ext.declarative.api.Base`

Describes relationship between genome, run, source BLAST database and query fragments.

Each genome and run combination can be assigned a single BLAST database for the comparisons

- fragpath path to fragmented genome (query in ANIb)
- dbpath path to source genome database (subject in ANIb)
- fragsizes JSONified dict of fragment sizes
- dbcmd command used to generate database

```python
blastdb_id
dbcmd
dbpath
fragpath
fragsizes
genome
genome_id
run
run_id
```
class pyani.pyani_orm.Comparison(**kwargs)
    Bases: sqlalchemy.ext.declarative.api.Base

    Describes a single pairwise comparison between two genomes.
    aln_length
    comparison_id
    cov_query
    cov_subject
    fragsize
    identity
    maxmatch
    program
    query
    query_id
    runs
    sim_errs
    subject
    subject_id
    version

class pyani.pyani_orm.Genome(**kwargs)
    Bases: sqlalchemy.ext.declarative.api.Base

    Describes an input genome for a pyani run.
    • genome_id primary key
    • genome_hash MD5 hash of input genome file (in path)
    • path path to FASTA genome file
    • length length of genome (total bases)
    • description genome description
blastdbs
description
genome_hash
genome_id
labels
length
path
query_comparisons
runs
subject_comparisons
class pyani.pyani_orm.Label(**kwargs)
    Bases: sqlalchemy.ext.declarative.api.Base
    Describes relationship between genome, run and genome label.
    Each genome and run combination can be assigned a single label
    class_label
    genome
    genome_id
    label
    label_id
    run
    run_id

class pyani.pyani_orm.LabelTuple
    Bases: tuple
    Label and Class for each file.
    class_label
        Alias for field number 1
    label
        Alias for field number 0

exception pyani.pyani_orm.PyaniORMException
    Bases: pyani.PyaniException
    Exception raised when ORM or database interaction fails.

class pyani.pyani_orm.Run(**kwargs)
    Bases: sqlalchemy.ext.declarative.api.Base
    Describes a single pyani run.
    blastdbs
    cmdline
    comparisons
    date
    df_alnlength
    df_coverage
    df_hadamard
    df_identity
    df_simerrors
    genomes
    labels
    method
    name
    run_id
**status**

`pyani.pyani_orm.add_run`(*session, method, cmdline, date, status, name*)

Create a new Run and add it to the session.

**Parameters**

- `session` – live SQLAlchemy session of pyani database
- `method` – string describing analysis run type
- `cmdline` – string describing pyani command-line for run
- `date` – datetime object describing analysis start time
- `status` – string describing status of analysis
- `name` – string - name given to the analysis run

Creates a new Run object with the passed parameters, and returns it.


→ List[T]

Add genomes for a run to the database.

**Parameters**

- `session` – live SQLAlchemy session of pyani database
- `run` – Run object describing the parent pyani run
- `indir` – path to the directory containing genomes
- `classpath` – path to the file containing class information for each genome
- `labelpath` – path to the file containing class information for each genome

This function expects a single directory (indir) containing all FASTA files for a run, and optional paths to plain text files that contain information on class and label strings for each genome.

If the genome already exists in the database, then a Genome object is recovered from the database. Otherwise, a new Genome object is created. All Genome objects will be associated with the passed Run object.

The session changes are committed once all genomes and labels are added to the database without error, as a single transaction.

`pyani.pyani_orm.create_db`(*dbpath: pathlib.Path*)

→ None

Create an empty pyani SQLite3 database at the passed path.

**Parameters**

- `dbpath` – path to pyani database

`pyani.pyani_orm.filter_existing_comparisons`(*session, run, comparisons, program, version, fragsize: Optional[int] = None, maxmatch: Optional[bool] = None*)

→ List[T]

Filter list of (Genome, Genome) comparisons for those not in the session db.

**Parameters**

- `session` – live SQLAlchemy session of pyani database
- `run` – Run object describing parent pyani run
- `comparisons` – list of (Genome, Genome) query vs subject comparisons
- `program` – program used for comparison
- `version` – version of program for comparison
- `fragsize` – fragment size for BLAST databases

5.12. pyani package 93
• **maxmatch** – maxmatch used with nucmer comparison

When passed a list of (Genome, Genome) comparisons as comparisons, check whether the comparison exists in
the database and, if so, associate it with the passed run. If not, then add the (Genome, Genome) pair to a list for
returning as the comparisons that still need to be run.

```python
pyani.pyani_orm.get_comparison_dict(session: Any) → Dict[Tuple, Any]
```

Return a dictionary of comparisons in the session database.

**Parameters**

- **session** – live SQLAlchemy session of pyani database

Returns Comparison objects, keyed by (_query_id, _subject_id, _program, _version, _fragsize, _maxmatch)
tuple

```python
pyani.pyani_orm.get_matrix_classes_for_run(session: Any, run_id: int) → Dict[str, List[T]]
```

Return dictionary of genome classes, keyed by row/column ID.

**Parameters**

- **session** – live SQLAlchemy session
- **run_id** – the Run.run_id value for matrices

The class labels should be valid for identity, coverage and other complete matrix results accessed via the .df_*
attributes of a run

Labels are returned keyed by the string of the genome ID, for compatibility with matplotlib.

```python
pyani.pyani_orm.get_matrix_labels_for_run(session: Any, run_id: int) → Dict[KT, VT]
```

Return dictionary of genome labels, keyed by row/column ID.

**Parameters**

- **session** – live SQLAlchemy session
- **run_id** – the Run.run_id value for matrices

The labels should be valid for identity, coverage and other complete matrix results accessed via the .df_* at-
tributes of a run.

Labels are returned keyed by the string of the genome ID, for compatibility with matplotlib.

```python
pyani.pyani_orm.get_session(dbpath: pathlib.Path) → Any
```

Connect to an existing pyani SQLite3 database and return a session.

**Parameters**

- **dbpath** – path to pyani database

```python
pyani.pyani_orm.update_comparison_matrices(session, run) → None
```

Update the Run table with summary matrices for the analysis.

**Parameters**

- **session** – active pyanidb session via ORM
- **run** – Run ORM object for the current ANIm run

**pyani.pyani_report module**

Module providing functions for presenting analysis/db output.

```python
pyani.pyani_report.colour_coverage(series: pandas.core.series.Series, threshold: float = 0.95, colour: str = '#FF2222') → List[str]
```

Highlight percent coverage over a threshold.

**Parameters**
- **series** –
- **threshold** – float, threshold for cell highlighting
- **colour** – str, hex colour for highlighted cells

```python
pyani.pyani_report.colour_identity(series: pandas.core.series.Series, threshold: float = 0.95, colour: str = '#FF2222') → List[str]
```
Highlight percentage identities over a threshold.

**Parameters**
- **series** –
- **threshold** – float, threshold for cell highlighting
- **colour** – str, hex colour for highlighted cells

```python
pyani.pyani_report.colour_numeric(val: float, threshold: float = 0.95, colour: str = '#FF2222') → str
```
Highlight numeric values over a threshold.

**Parameters**
- **val** –
- **threshold** – float, threshold for cell highlighting
- **colour** – str, hex colour for highlighted cell

```python
pyani.pyani_report.colour_rows(series: pandas.core.series.Series, even_colour: str = '#DDECF5', odd_colour: str = '#6CB6E4') → List[str]
```
Return alternating colours for rows in a dataframe.

**Parameters**
- **series** – pd.Series
- **even_colour** – str, hex colour for even rows
- **odd_colour** – str, hex colour for odd rows

```python
pyani.pyani_report.header_font() → Dict[str, Any]
```
Return header HTML font style.

```python
pyani.pyani_report.hover_highlight(hover_colour: str = '#FFFF99') → Dict[str, Any]
```
Return HTML style to colour dataframe row when hovering.

**Parameters**
- **hover_colour** – str, hex colour for hover highlight

```python
pyani.pyani_report.table_padding() → Dict[str, Any]
```
Return HTML for table cell padding.

```python
```
Write database result table to output file in named format.

**Parameters**
- **dfm** – pd.DataFrame
- **path** – Path to output file
- **formats** – tuple of str, output file formats
- **index** – Boolean
• **show_index** – Boolean
• **colour_num** – use colours for values in HTML output

colours are used for identity/coverage tables

```python
```

Add CSS styling to a dataframe and write as HTML.

**Parameters**

- **path** – path to write output file
- **dfm** – dataframe to be written out
- **index** – column to be set as index (if necessary)

```python
pyani.pyani_report.write_to_stdout(stem: str, dfm: pandas.core.frame.DataFrame, show_index: bool = False, line_width: float = None) -> None
```

Write dataframe in tab-separated form to STDOUT.

**Parameters**

- **stem** – str
- **dfm** – pd.DataFrame
- **show_index** – Boolean, include index in output table
- **line_width** –

---

**pyani.pyani_tools module**

Code to support pyani.

```python
class pyani.pyani_tools.ANIResults(labels: List[str], mode: str)
    Bases: object

Holds ANI dataframe results.

    add_coverage(qname: str, sname: str, qcover: float, scover: Optional[float] = None) -> None
    Add percentage coverage values to self.alignment_coverage.

    Parameters
    • qname –
    • sname –
    • value –
    • sym –

    add_pid(qname: str, sname: str, value: float, sym: bool = True) -> None
    Add a percentage identity value to self.percentage_identity.

    Parameters
    • qname –
    • sname –
    • value –
```

---

Chapter 5. Reporting problems and requesting improvements
• sym

**add_sim_errors** (*qname: str, sname: str, value: float, sym: bool = True*) → None

Add a similarity error value to self.similarity_errors.

**Parameters**

• *qname* –
• *sname* –
• *value* –
• *sym* –

**add_tot_length** (*qname: str, sname: str, value: float, sym: bool = True*) → None

Add a total length value to self.alignment_lengths.

**Parameters**

• *qname* –
• *sname* –
• *value* –
• *sym* –

data

Return list of (dataframe, filestem) tuples.

**hadamard**

Return Hadamard matrix (identity * coverage).

class pyani.pyani_tools.BLASTcmds (funcs: pyani.pyani_tools.BLASTfunctions, exes: pyani.pyani_tools.BLASTexes, prefix: str, outdir: pathlib.Path)

**Bases:** object

Class for construction of BLASTN and database formatting commands.

**build_blast_cmd** (*fname: pathlib.Path, dbname: pathlib.Path*)

Return BLASTN command.

**Parameters**

• *fname* – Path to query file
• *dbname* – Path to database

**build_db_cmd** (*fname: pathlib.Path*) → str

Return database format/build command.

**Parameters**

• *fname* –

**get_db_name** (*fname: pathlib.Path*) → str

Return database filename.

**Parameters**

• *fname* –

class pyani.pyani_tools.BLASTexes

**Bases:** tuple

Convenience structure to hold BLAST executables.

**blast_exe**

Alias for field number 1
```python
format_exe
    Alias for field number 0

class pyani.pyani_tools.BLASTfunctions
    Bases: tuple
    Convenience structure to hold BLAST functions.
    blastn_func
        Alias for field number 1
    db_func
        Alias for field number 0

class pyani.pyani_tools.Dependencies
    Bases: tuple
    Convenience struct for third-party dependency presence.
    blast
        Alias for field number 0
    legacy_blast
        Alias for field number 1
    mummer
        Alias for field number 2

class pyani.pyani_tools.MatrixData
    Bases: tuple
    Convenience struct for matrix data returned by ORM.
    data
        Alias for field number 1
    graphic_args
        Alias for field number 2
    name
        Alias for field number 0

def get_genome_length(filename: pathlib.Path) → int
    Return total length of all sequences in a FASTA file.

    Parameters
        filename -- path to FASTA file

def get_labels(filename: pathlib.Path, logger: logging.Logger = None) → Dict[KT, VT]
    Return dictionary of alternative sequence labels, or None.

    Parameters
        • filename -- path to file containing tab-separated table of labels
        • logger -- logging object

    Input files should be formatted as <hash><key><label>, one pair per line.

def has_dependencies() → pyani.pyani_tools.Dependencies
    Return NamedTuple indicating if 3rd dependencies are available.

def label_results_matrix(matrix: pandas.core.frame.DataFrame, labels: Dict[KT, VT]) → pandas.core.frame.DataFrame
    Return results matrix dataframe with labels.
```
Parameters

- **matrix** – results dataframe deriving from Run object
- **labels** – dictionary of genome labels labels must be keyed by index/col values from matrix

Applies the labels from the dictionary to the dataframe in matrix, and returns the result.

```python
pyani.pyani_tools.make_asm_dict(taxon_ids: Iterable[str], retries: int) → Dict[KT, VT]
```

Return dict of assembly UIDs, keyed by each passed taxon ID.

Parameters

- **taxon_ids** – Iterable of NCBI taxonomy IDs
- **retries** – Number of Entrez retry attempts

**pyani.run_multiprocessing module**

Code to run a set of command-line jobs using multiprocessing.

For parallelisation on multi-core desktop/laptop systems, etc. we use Python’s multiprocessing module to distribute command-line jobs.

```python
pyani.run_multiprocessing.multiprocessing_run(cmdlines: List[T], workers: Optional[int] = None) → int
```

Distributes passed command-line jobs using multiprocessing.

Parameters

- **cmdlines** – iterable, command line strings
- **workers** – int, number of workers to use for multiprocessing

Returns the sum of exit codes from each job that was run. If all goes well, this should be 0. Anything else and the calling function should act accordingly.

```python
```

Create list of jobsets at different depths of dependency tree.

Parameters

- **job**
- **cmdsets**
- **depth**

This is a recursive function (is there something quicker in the itertools module?) that descends each ‘root’ job in turn, populating each

```python
pyani.run_multiprocessing.run_dependency_graph(jobgraph, workers: Optional[int] = None, logger: Optional[logging.Logger] = None) → int
```

Create and run pools of jobs based on the passed jobgraph.

Parameters

- **jobgraph** – list of jobs, which may have dependencies.
- **workers** – int, number of workers to use with multiprocessing
- **logger** – a logger module logger (optional)
The strategy here is to loop over each job in the list of jobs (jobgraph), and create/populate a series of Sets of commands, to be run in reverse order with multiprocessing_run as asynchronous pools.

**pyani.run_sge module**

Code to run a set of command-line jobs using SGE/Grid Engine.

For parallelisation on multi-node system, we use some custom code to submit jobs.

```python
pyani.run_sge.build_and_submit_jobs(root_dir: pathlib.Path, jobs: Iterable[T_co], sgeargs: Optional[str] = None) → None
```

Submit passed iterable of Job objects to SGE.

**Parameters**

- `root_dir` – root directory for SGE and job output
- `jobs` – list of Job objects, describing each job to be submitted
- `sgeargs` – str, additional arguments to qsub

This places SGE's output in the passed root directory

```python
pyani.run_sge.build_directories(root_dir: pathlib.Path) → None
```

Construct the subdirectories output, stderr, stdout, and jobs.

**Parameters**

- `root_dir` – path of root directory in which to place output

Subdirectories are created in the passed root directory. These subdirectories have the following roles:

- `jobs` Stores the scripts for each job
- `stderr` Stores the stderr output from SGE
- `stdout` Stores the stdout output from SGE
- `output` Stores output (if the scripts place the output here)

- `root_dir` Path to the top-level directory for creation of subdirectories

```python
pyani.run_sge.build_job_scripts(root_dir: pathlib.Path, jobs: List[T]) → None
```

Construct script for each passed Job in the jobs iterable.

**Parameters**

- `root_dir` – Path to output directory
- `jobs` –

```python
pyani.run_sge.build_joblist(jobgraph) → List[T]
```

Return a list of jobs, from a passed jobgraph.

**Parameters**

- `jobgraph` –

```python
pyani.run_sge.compile_jobgroups_from_joblist(joblist: List[T], jgprefix: str, sgegrousize: int) → List[T]
```

Return list of jobgroups, rather than list of jobs.

**Parameters**

- `joblist` –
- `jgprefix` – str, prefix for SGE jobgroup
- `sgegrousize` – int, number of jobs in each SGE jobgroup

```python
pyani.run_sge.extract_submittable_jobs(waiting: List[T]) → List[T]
```

Obtain list of jobs that are able to be submitted from pending list.

**Parameters**

- `waiting` – list of Job objects
Create set of jobs reflecting dependency tree.

Parameters
- job
- jobset
- depth

The set contains jobs at different depths of the dependency tree, retaining dependencies as strings, not Jobs.

Create and runs SGE scripts for jobs based on passed jobgraph.

Parameters
- jobgraph – list of jobs, which may have dependencies.
- verbose – flag for multiprocessing verbosity
- logger – a logger module logger (optional)
- jgprefix – a prefix for the submitted jobs, in the scheduler
- sgegroupsize – the maximum size for an array job submission
- sgeargs – additional arguments to qsub

The strategy here is to loop over each job in the dependency graph and, because we expect a single main delta-filter (wrapped) job, with a single nucmer dependency for each analysis, we can split the dependency graph into two lists of corresponding jobs, and run the corresponding nucmer jobs before the delta-filter jobs.

pyani.run_sge.split_seq (iterable: Iterable[T_co], size: int) → Generator[T_co, T_contra, V_co]
Split a passed iterable into chunks of a given size.

Parameters
- iterable – iterable
- size – int, number of items to return in each chunk

pyani.run_sge.submit_jobs (root_dir: pathlib.Path, jobs: Iterable[T_co], sgeargs: Optional[str] = None) → None
Submit passed jobs to SGE server with passed directory as root.

Parameters
- root_dir – path to output directory
- jobs – list of Job objects
- sgeargs – str, additional arguments for qsub

pyani.run_sge.submit_safe_jobs (root_dir: pathlib.Path, jobs: Iterable[T_co], sgeargs: Optional[str] = None) → None
Submit passed list of jobs to SGE server with dir as root for output.

Parameters
- root_dir – path to output directory
- jobs – iterable of Job objects
- sgeargs – str, additional arguments for qsub
**pyani.tetra module**

Code to implement the TETRA average nucleotide identity method.

Provides functions for calculation of TETRA as described in:


and


**pyani.tetra.calculate_correlations** *(tetra_z: Dict[str, Dict[str, float]]) → pandas.core.frame.DataFrame*

Return dataframe of Pearson correlation coefficients.

**Parameters tetra_z** – dict, Z-scores, keyed by sequence ID

Calculates Pearson correlation coefficient from Z scores for each tetranucleotide. This is done longhand here, which is fast enough, but for robustness we might want to do something else... (TODO).

Note that we report a correlation by this method, rather than a percentage identity.

**pyani.tetra.calculate_tetra_zscore** *(filename: pathlib.Path) → Dict[str, float]*

Return TETRA Z-score for the sequence in the passed file.

**Parameters filename** – path to sequence file

Calculates mono-, di-, tri- and tetranucleotide frequencies for each sequence, on each strand, and follows Teeling et al. (2004) in calculating a corresponding Z-score for each observed tetranucleotide frequency, dependent on the mono-, di- and tri- nucleotide frequencies for that input sequence.

**pyani.tetra.calculate_tetra_zscores** *(infilenames: Iterable[T_co]) → Dict[str, Dict[str, float]]*

Return dictionary of TETRA Z-scores for each input file.

**Parameters infilenames** – iterable of paths to input sequence files

**pyani.tetra.tetra_clean** *(instr: str) → bool*

Return True if string contains only unambiguous IUPAC nucleotide symbols.

**Parameters instr** – str, nucleotide sequence

We are assuming that a low frequency of IUPAC ambiguity symbols doesn’t affect our calculation.

### 5.13 Indices and tables

- genindex
- modindex
- search
Python Module Index

p
pyani, 49
pyani.anib, 74
pyani.anim, 78
pyani.blast, 80
pyani.download, 80
pyani.nucmer, 85
pyani.pyani_classify, 86
pyani.pyani_config, 87
pyani.pyani_files, 87
pyani.pyani_graphics, 89
pyani.pyani_jobs, 89
pyani.pyani_orm, 90
pyani.pyani_report, 94
pyani.pyani_tools, 96
pyani.run_multiprocessing, 99
pyani.run_sge, 100
pyani.scripts, 49
pyani.scripts.average_nucleotide_identity, 62
pyani.scripts.delta_filter_wrapper, 68
pyani.scripts.genbank_get_genomes_by_taxon, 69
pyani.scripts.logger, 72
pyani.scripts.parsers, 49
pyani.scripts.parsers.anib_parser, 50
pyani.scripts.parsers.aniblastall_parser, 50
pyani.scripts.parsers.anib_parser, 50
pyani.scripts.parsers.classify_parser, 51
pyani.scripts.parsers.common_parser, 51
pyani.scripts.parsers.createdb_parser, 51
pyani.scripts.parsers.download_parser, 52
pyani.scripts.parsers.index_parser, 52
pyani.scripts.parsers.plot_parser, 52
pyani.scripts.parsers.report_parser, 53
pyani.scripts.parsers.run_common_parser, 53
pyani.scripts.parsers.scheduling_parser, 53
pyani.scripts.pyani_script, 72
pyani.scripts.subcommands, 53
pyani.scripts.subcommands.subcmd_anib, 54
pyani.scripts.subcommands.subcmd_aniblastall, 55
pyani.scripts.subcommands.subcmd_anim, 55
pyani.scripts.subcommands.subcmd_classify, 58
pyani.scripts.subcommands.subcmd_createdb, 59
pyani.scripts.subcommands.subcmd_download, 59
pyani.scripts.subcommands.subcmd_index, 60
pyani.scripts.subcommands.subcmd_plot, 60
pyani.scripts.subcommands.subcmd_report, 61
pyani.scripts.tools, 72
pyani.tetra, 102
Index

A

add_alignment() (pyani.nucmer.DeltaComparison method), 85
add_coverage() (pyani.pyani_tools.ANIResults method), 96
add_dependency() (pyani.pyani_jobs.Job method), 89
add_dependency() (pyani.pyani_jobs.JobGroup method), 90
add_pid() (pyani.pyani_tools.ANIResults method), 96
add_run() (in module pyani.pyani_orm), 93
add_run_genomes() (in module pyani.pyani_orm), 93
add_sim_errors() (pyani.pyani_tools.ANIResults method), 97
add_tot_length() (pyani.pyani_tools.ANIResults method), 97
all_components_k_complete() (in module pyani.pyani_classify), 86
all_k_complete (pyani.pyani_classify.Cliquesinfo attribute), 86
aln_length (pyani.pyani_orm.Comparison attribute), 91
aln_length (pyani.pyani_tools.subcmd_anim.ComparisonResult attribute), 55
analyse_cliques() (in module pyani.pyani_classify), 86
ANIResults (class in pyani.pyani_tools), 96
asm_ids (pyani.download.ASMIDs attribute), 80
ASMIDs (class in pyani.download), 80

B

blast (pyani.pyani_tools.Dependencies attribute), 98
blast_exe (pyani.pyani_tools.BLASTexes attribute), 97
BLASTcmds (class in pyani.pyani_tools), 97
BlastDB (class in pyani.pyani_orm), 90
blastdb_id (pyani.pyani_orm.BlastDB attribute), 90
blastdbs (pyani.pyani_orm.Genome attribute), 91

build() (in module pyani.scripts.parsers.anib_parser), 50
build() (in module pyani.scripts.parsers.aniblastall_parser), 50
build() (in module pyani.scripts.parsers.anim_parser), 50
build() (in module pyani.scripts.parsers.classify_parser), 51
build() (in module pyani.scripts.parsers.createdb_parser), 51
build() (in module pyani.scripts.parsers.download_parser), 52
build() (in module pyani.scripts.parsers.index_parser), 52
build() (in module pyani.scripts.parsers.plot_parser), 52
build() (in module pyani.scripts.parsers.report_parser), 53
build() (in module pyani.scripts.parsers.run_common_parser), 53
build() (in module pyani.scripts.parsers.scheduling_parser), 53
build_and_submit_jobs() (in module pyani.run_sge), 100
build_blast_cmd() (pyani.pyani_tools.BLASTcmds method), 97
build_db_cmd() (pyani.pyani_tools.BLASTcmds method), 97
build_db_jobs() (in module pyani.anib), 74
<table>
<thead>
<tr>
<th>Module</th>
<th>Function/Method</th>
<th>Line</th>
</tr>
</thead>
<tbody>
<tr>
<td>pyani.scripts.average_nucleotide_identity</td>
<td>get_ncbi_asm() (in module pyani.scripts.genbank_get_genomes_by_taxon)</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>get_ncbi_classification() (in module pyani.download)</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>get_ncbi_esummary() (in module pyani.download)</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>graph (pyani.scripts.subcommands.subcmd_classify.SubgraphData attribute)</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>graphic_args (pyani.pyani_tools.MatrixData attribute)</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>hadamard (pyani.pyani_toolsANIResults attribute)</td>
<td>97</td>
</tr>
<tr>
<td></td>
<td>has_dependencies() (in module pyani.pyani_tools)</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>Hashstatus (class in pyani.download)</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td>header_font() (in module pyani.pyani_report)</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>headers (pyani.scripts.subcommands.subcmd_report.ReportParams attribute)</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>hover_highlight() (in module pyani.pyani_report)</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>identity (pyani.pyani_orm.Comparison attribute)</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>interval (pyani.scripts.subcommands.subcmd_classify.SubgraphData attribute)</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>Job (class in pyani.pyani_jobs)</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>job (pyani.scripts.subcommands.subcmd_anim.ComparisonJob attribute)</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>JobGroup (class in pyani.pyani_jobs)</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>k_complete_component_status() (in module pyani.pyani_classify)</td>
<td>87</td>
</tr>
<tr>
<td></td>
<td>labels (pyani.pyani_orm.Run attribute)</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>LabelTuple (class in pyani.pyani_orm)</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>last_exception() (in module pyani.download)</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td>last_exception() (in module pyani.scripts.average_nucleotide_identity)</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>legacy_blast (pyani.pyani_tools.Dependencies attribute)</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>length (pyani.pyani_orm.Genome attribute)</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>load_classes_labels() (in module pyani.pyani_files)</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>logreport_downloaded() (in module pyani.download.Hashstatus attribute)</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td>make_asm_dict() (in module pyani.pyani_tools)</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>make_asm_dict() (in module pyani.scripts.tools)</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td>make_blastcmd_builder() (in module pyani.anib)</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td>make_job_graph() (in module pyani.anib)</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td>make_outdir() (in module pyani.scripts.tools)</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td>make_outdirs() (in module pyani.scriptsдавнные.get_genomes_by_taxon), 65</td>
<td></td>
</tr>
<tr>
<td></td>
<td>make_sequence_fragments() (in module pyani.scripts.average_nucleotide_identity)</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>MatrixData (class in pyani.pyani_tools)</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>maxmatch (pyani.pyani_orm.Comparison attribute)</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>maxmatch (pyani.scripts.subcommands.subcmd_anim.ProgParams attribute)</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>metadata (pyani.nucmer.DeltaData attribute)</td>
<td>85</td>
</tr>
<tr>
<td></td>
<td>method (pyani.pyani_orm.Run attribute)</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>method (pyani.scripts.subcommands.subcmd_anim.RunData attribute)</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>multiprocessing_run() (in module pyani.run_multiprocessing)</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>mummer (pyani.pyani_tools.Dependencies attribute)</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>n_nodes (pyani.pyani_classify.Cliquesinfo attribute)</td>
<td>86</td>
</tr>
<tr>
<td></td>
<td>n_subgraphs (pyani.pyani_classify.Cliquesinfo attribute)</td>
<td>86</td>
</tr>
<tr>
<td></td>
<td>name (pyani.pyani_orm.Run attribute)</td>
<td>86</td>
</tr>
</tbody>
</table>
name (pyani.scripts.subcommands.subcmd_anib.RunData
attribute), 56
name (pyani.scripts.subcommands.subcmd_report.ReportResult
attribute), 56
NCBIException, 73
NCBIException, 69, 81
nucmercmd (pyani.scripts.subcommands.subcmd_anib.ComparisonJob
attribute), 55
Organism (pyani.download.Classification attribute), 81
outfile (pyani.scripts.subcommands.subcmd_anib.ComparisonJob
attribute), 55

P
Params (class in pyani.pyani_graphics), 89
params_mpl () (in module pyani.pyani_config), 87
parse_api_key () (in module pyani.download), 84
parse_blast_tab () (in module pyani.anib), 77
parse_blasttab () (in module pyani.blast), 80
parse_cmdline () (in module
pyani.scripts.average_nucleotide_identity), 66
parse_cmdline () (in module
pyani.scripts.genbank_get_genomes_by_taxon), 71
parse_cmdline () (in module pyani.scripts.parsers), 49
parse_delta () (in module pyani.anib), 79
passed (pyani.download.Hashstatus attribute), 81
path (pyani.pyani.orm.Genome attribute), 91
pid (pyani.scripts.subcommands.subcmd_anib.ComparisonJob
attribute), 55
populate_cmdsets () (in module
pyani.run_multiprocessing), 99
populate_jobset () (in module pyani.run_sge), 100
process_arguments () (in module
pyani.scripts.average_nucleotide_identity),
66
process_blast () (in module pyani.anib), 77
process_deltadir () (in module pyani.anib), 80
process_formats () (in module
pyani.scripts.subcommands.subcmd_report),
61
ProgData (class in pyani.scripts.subcommands.subcmd_anib),
56
ProgParams (class in
pyani.scripts.subcommands.subcmd_anib),
56
program (pyani.nucmer.DeltaData attribute), 85
program (pyani.pyani.orm.Comparison attribute), 91
program (pyani.scripts.subcommands.subcmd_anib.ProgData
attribute), 56
pyani (module), 49
pyani.anib (module), 74
statement (pyani.scripts.subcommands.subcmd_report.ReportParams attribute), 61
status (pyani.pyani_orm.Run attribute), 92
strain (pyani.download.Classification attribute), 81
subcmd_anib() (in module pyani.scripts.subcommands.subcmd_anib), 54
subcmd_aniblastall() (in module pyani.scripts.subcommands.subcmd_aniblastall), 55
subcmd_anim() (in module pyani.scripts.subcommands.subcmd_anim), 57
subcmd_classify() (in module pyani.scripts.subcommands.subcmd_classify), 58
subcmd_createdb() (in module pyani.scripts.subcommands.subcmd_createdb), 59
subcmd_download() (in module pyani.scripts.subcommands.subcmd_download), 59
subcmd_index() (in module pyani.scripts.subcommands.subcmd_index), 60
subcmd_plot() (in module pyani.scripts.subcommands.subcmd_plot), 60
subcmd_report() (in module pyani.scripts.subcommands.subcmd_report), 62
SubgraphData (class in pyani.scripts.subcommands.subcmd_classify), 58
subject (pyani.pyani_orm.Comparison attribute), 91
subject (pyani.scripts.subcommands.subcmd_anim.ComparisonJob attribute), 55
subject_comparisons (pyani.pyani_orm.Genome attribute), 91
subject_id (pyani.pyani_orm.Comparison attribute), 91
submit_jobs() (in module pyani.run_sge), 101
submit_safe_jobs() (in module pyani.run_sge), 101
subsample_input() (in module pyani.scripts.average_nucleotide_identity), 66
suffix (pyani.scripts.tools.DLFileData attribute), 73
T

Table_padding() (in module pyani.pyani_report), 95

Test_class_label_paths() (in module pyani.scripts.average_nucleotide_identity), 67

U

tetra_clean() (in module pyani.tetra), 102
trimmed_graph_sequence() (in module pyani.scripts.subcommands.subcmd_classify), 58

U

unified_anib() (in module pyani.scripts.average_nucleotide_identity), 67
update_comparison_matrices() (in module pyani.pyani_orm), 94
update_comparison_results() (in module pyani.scripts.subcommands.subcmd_anim), 57

V

vdiff (pyani.pyani_graphics.Params attribute), 89
version (pyani.pyani_orm.Comparison attribute), 91
version (pyani.scripts.subcommands.subcmd_anim.ProgData attribute), 56

W

wait() (pyani.pyani_jobs.Job method), 89
wait() (pyani.pyani_jobs.JobGroup method), 90
write() (in module pyani.scripts.average_nucleotide_identity), 68
write_contigs() (in module pyani.scripts.genbank_get_genomes_by_taxon), 72
write_dbtable() (in module pyani.pyani_report), 95
write_distribution() (in module pyani.scripts.subcommands.subcmd_plot), 60
write_heatmap() (in module pyani.scripts.subcommands.subcmd_plot), 60
write_run_heatmaps() (in module pyani.scripts.subcommands.subcmd_plot), 61
write_styled_html() (in module pyani.pyani_report), 96
write_to_stdout() (in module pyani.pyani_report), 96