Molecular Microbiology 2024

Bioinformatics Protocol 8

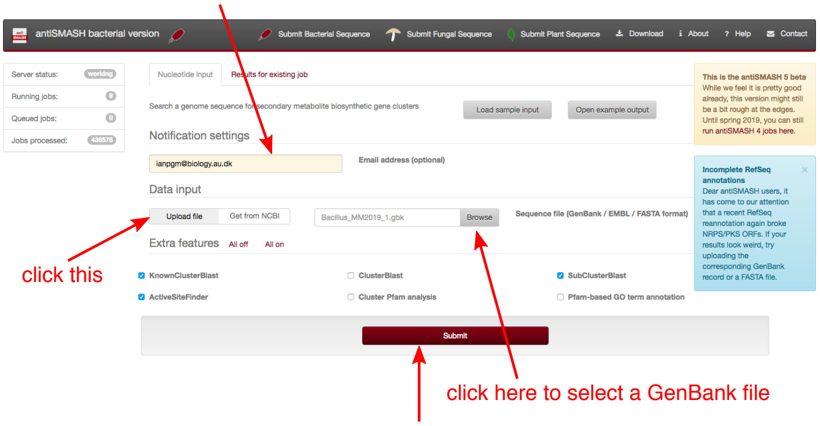
Post-annotation analysis: Pathway analysis and other web services, running BLAST, running HMMer

This tutorial covers the use of various web-based services that will give you deeper insights into your genome. Step 2 (KofamKoala) will be useful for everyone, while steps 3—6 are optional, depending on your interests. Furthermore, it will cover searching your genome for a gene with a specific function using either **BLAST** or **HMMer**.

Service	Purpose
KofamKOALA	Metabolic Pathways
antiSMASH	Secondary Metabolites
CARD Resistance Gene Identifier	Antibiotic Resistance Genes
CRISPRloci	CRISPR-Cas phage defense systems
Phaster	Prophages

1. Download the **Prokka** output directory to your hard drive.
2. Open a web browser and visit https://www.genome.jp/tools/kofamkoala/. Under *or upload a sequence file* click *Browse...*, navigate to and select your `.faa` file on your hard drive. Type your email address under *Enter your email address* then click on *Request for email confirmation.* You will shortly receive an email containing two links - one to submit your job and the other to cancel the job. Click on the link to submit your job. While BlastKOALA is running proceed with the rest of the protocol.
3. Open a web browser and visit https://antismash.secondarymetabolites.org/#!/start. Enter you Email address in the required field. Select *Upload File*, then click *Browse* and navigate your hard drive to the genbank file output from your prokka annotation (`.gbk` file). Leave other tickboxes as default and click on *Submit.* Continue with the rest of the protocol while it's running.

enter your email address here



Server status: working
Running jobs: 9
Queued jobs: 0
Jobs processed: 436578

Search a genome sequence for secondary metabolite biosynthetic gene clusters

Notification settings

Email address (optional): ianpgm@biology.au.dk

Data input

Sequence file (GenBank / EMBL / FASTA format): Bacillus_MM2019_1.gbk

Extra features

☑ KnownClusterBlast
☐ ClusterBlast
☑ SubClusterBlast
☑ ActiveSiteFinder
☐ Cluster Pfam analysis
☐ Pfam-based GO term annotation

This is the antiSMASH 5 beta
While we feel it is pretty good already, this version might still be a bit rough at the edges. Until spring 2019, you can still run antiSMASH 4 jobs here.

Incomplete RefSeq annotations
Dear antiSMASH users, it has come to our attention that a recent RefSeq reannotation again broke NRPS/PKS ORFs. If your results look weird, try uploading the corresponding GenBank record or a FASTA file.

click this

click here to select a GenBank file

click here to submit

4. Open a web browser and visit https://card.mcmaster.ca/analyze/rgi. Under “Upload FASTA Sequence Files” click “Browse. . . ” and navigate to the file containing your scaffolds in fast format, the `.fna` file. Click *Submit*. When it finishes running you will see a list of the antibiotic resistance genes identified in your genome.

5. Open a web browser and visit https://rna.informatik.uni-freiburg.de/CRISPRloci/. Click *Browse. . .* and navigate to your genome assembly file to upload. Leave other options as defaults and click on *START*. `CRISPRloci` will now start annotating CRISPR-Cas systems on your genome (it will take about 15 minutes).

6. Open a web browser and visit http://phaster.ca/. Under “Select an input type”, click on *CHOOSE FILE* and select the fna file containing all scaffolds (`.fna`) from your prokka annotation. Tick the box that says it’s a fasta file with multiple contigs. Then click *SUBMIT* and wait.

7. Log into the dnaseq1a server and copy the fasta amino-acid file from your prokka output directory to your `genome` directory.

   ```
   cp genome/assembly/prokka_annotation/output_filename_prefix.faa genome/
   ```

   Modify the `assembly/prokka_annotation` directory and `output_filename_prefix.faa` in the above command to match the names of your prokka output directory and fasta amino-acid file.

8. Change into the genome directory and make a BLAST database from your genome.

   ```
   cd genome
   makeblastdb -in output_filename_prefix.faa -dbtype prot
   ```

   - `-in output_filename_prefix.faa` - This specifies the names of the fasta file containing the sequences to be included in the BLAST database. You should change this to match the name of your fasta amino-acid file.

    - `-dbtype prot` - The `dbtype` parameter specifies whether the sequences are proteins (`prot`) or nucleotides (`nucl`) - in this case it's proteins.

9. Make a query sequence file by copying and pasting the following FASTA-formatted sequence to a blank document in **Sublime Text** or **VS Code**. This sequence is ribosomal protein S2 from E. coli K-12, but you could substitute this for any other sequence or set of sequences you are interesting in searching for in your genome. Save it as `ribosomal_protein_S2.faa` and upload the query file to your server and place it in the `genome` directory. In **Sublime Text** make sure your line endings are set to Unix (View->Line Endings->Unix) before you save.

   ```
   >2688079486  SSU ribosomal protein S2P [Escherichia coli K-12 MG1655]
   MATVSMRDMLKAGVHFGHQTRYWNPKMKPFIFGARNKVHIINLEKTVPMF
   NEALAELNKIASRKGKILFVGTKRAASEAVKDAALSCDQFFVNHRWLGGM
   LTNWKTVRQSIKRLKDLETQSQDGTFDKLTKKEALMRTRELEKLENSLGG
   IKDMGGLPDALFVIDADHEHIAIKEANNLGIPVFAIVDTNSDPDGVDFVI
   PGNDDAIRAVTLYLGAVAATVREGRSQDLASQAEESFVEAE
   ```

10. Back in **Terminal**, use `blastp` and the query sequence you just made to search the database you made in step 4.

    ```
    blastp \
    -query ribosomal_protein_S2.faa \
    -db output_filename_prefix.faa \
    -num_threads 3 \
    -out blast_ribosomal_protein_S2_vs_output_filename_prefix
    ```

    - `blastp` - This is the name of the program for BLASTing protein sequences against protein databases. Use `blastn` for nucleotide sequences, and `blastx` to BLAST nucleotide sequences against a protein database in all 6 reading frames.
    - `-query ribosomal_protein_S2.faa` - This specifies your query sequence. In this case we are looking for the sequence in the file from step 5 `ribosomal_protein_S2.faa`, but later on in the class you may use different query sequences.
    - `-db output_filename_prefix.faa` - `db` stands for "database", and this is where you specify the database to be used (in this case, your genome). You need to modify this to the database filename you used in step 4.
    - `-num_threads 3` - This specifies the number of threads BLAST may use to do the search. There are 4 CPUs on your server, so set this to 3.
    - `-out blast_ribosomal_protein_S2_vs_output_filename_prefix` - This specifies your output file name. I find it useful to use the format `query_name_vs_database_name` when naming BLAST output files. You should modify this name to something that makes sense for you.

11. Download the BLAST output file `blast_ribosomal_protein_S2_vs_output_filename_prefix` to your computer and open it in **Sublime Text** or **VS Code**. Do the hits make sense? Do they agree with the Prokka annotation? Note in particular the e-values, bitscores, identities, and positives. Look at the

sequence alignment of the query and the database sequence - what's the difference between "identities" and "positives"?

12. Open a web browser and go to (https://www.ebi.ac.uk/interpro/search/text/)[https://www.ebi.ac.uk/inte In the search box type `Ribosomal_S2` and click Search. You should see a list of matches including one from the PFAM database -click on *Ribosomal Protein S2* to be taken to the page for this protein family model. Click on *Curation* in the left-hand menu and then on *download the raw HMM for this family* at the bottom of the screen. This will download a hidden markov model profile called `PF00318.hmm.gz` to your computer. Upload this file to the `genome` directory of your server.

13. Back in **Terminal**, use `hmmpress` to prepare the hidden markov model for searching your genome.

```
gunzip PF00318.hmm.gz
hmmpress PF00318.hmm
```

14. Use `hmmsearch` to search your genome for any protein sequences that match the `Ribosomal_S2.hmm` profile.

```
hmmsearch \
-o  hmmer_Ribosomal_S2_vs_output_filename_prefix \
--cpu 3 \
Ribosomal_S2.hmm \
output_filename_prefix.faa
```

- `hmmsearch` - This specifies the `hmmsearch` program from the suite of HMMer programs
- `-o  hmmer_Ribosomal_S2_vs_output_filename_prefix` - This specifies the name of the output file. It's useful to use the format `profile_name_vs_genome_name` when naming HMMer output files. You should modify this name to something that makes sense for you.
- `--cpu 3` - This specifies the number of CPUs (3 for the servers we're using).
- `Ribosomal_S2.hmm` - This specifies the name of the HMM profile you are searching for.
- `output_filename_prefix.faa` - This specifies the name of the fasta amino-acid file for your genome of interest. You should change this to match the right filename.

15. Download the HMMer output file `hmmer_Ribosomal_S2_vs_output_filename_prefix` and open it in **Sublime Text** or **VS Code**. Did HMMer identify the same protein sequence as BLAST did? Which one had more potential false positives? Reflect on the various advantages and disadvantages of using HMMer and BLAST to find specific protein sequences in a genome.

16. Try steps 7—10 or 11—15 with different proteins of interest. You can search for different HMMs on https://www.ebi.ac.uk/interpro/search/text/ and find query sequences for BLAST by searching for functions on IMG or the NCBI protein database. You can also place more than one sequence or HMM in your query file (just put the text for each query one after the other in the `.hmm` file or the `.fasta` query file) to search for more than one HMM or protein sequence at once.

17. When KofamKOALA has completed (around one hour), you will receive an email with a link. Click on the link and you will see a summary of your results. First click *Download text file* to obtain the KO (KEGG ontology) annotation of your genome (called `result.txt`).

18. Click on *KEGG Mapper* to see the pathway maps for your genome. Here you will see a long list of different pathways and protein sets that you can click on to examine more closely. In pathway maps, genes encoding enzymes highlighted in green are present in your genome, while genes encoding enzymes without highlighting have not been identified.