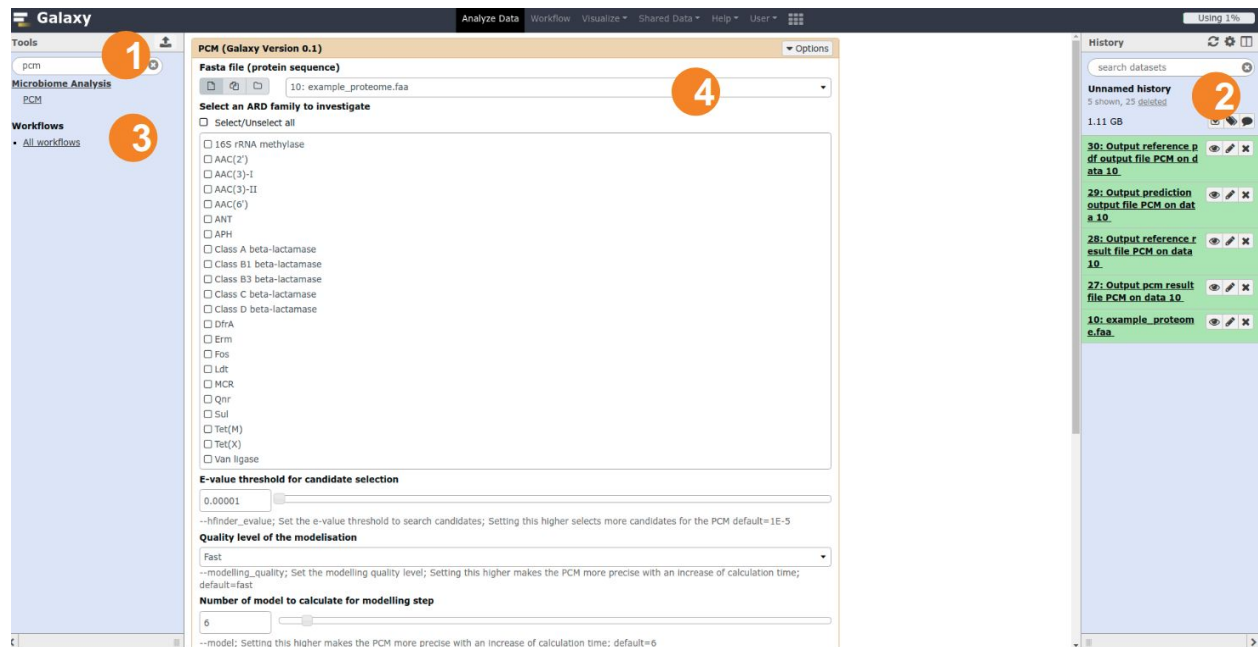


Usage of PCM in galaxy

PCM is deployed on galaxy.pasteur.fr in section microbiome_analysis/PCM (3) as follow:



First, a fasta file with protein sequence need to be loaded in an history to be available for the workflow (1).

The file must be in the following format:

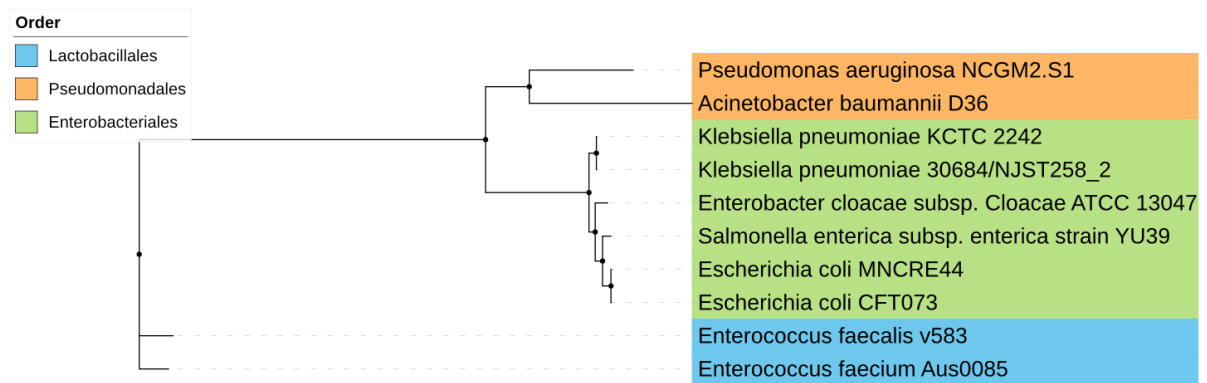
```
>ID1
MNTFGQIHNNMPYLFLAFIMNFYDQFNNSISGQEMCYEVESI
FNNHQVDIIIGAPAAAFKPLELQKGLGTKGAIVNYPILQVTGNI
>ID2
MNTFGQIHNNMPYLFLAFIMNFYDQFNNSISGQEMCYEVESI
FNNHQVDIIIGAPAAAFKPLELQKGLGTKGAIVNYPILQVTGNI
```

The ID is >name (without any space or points).

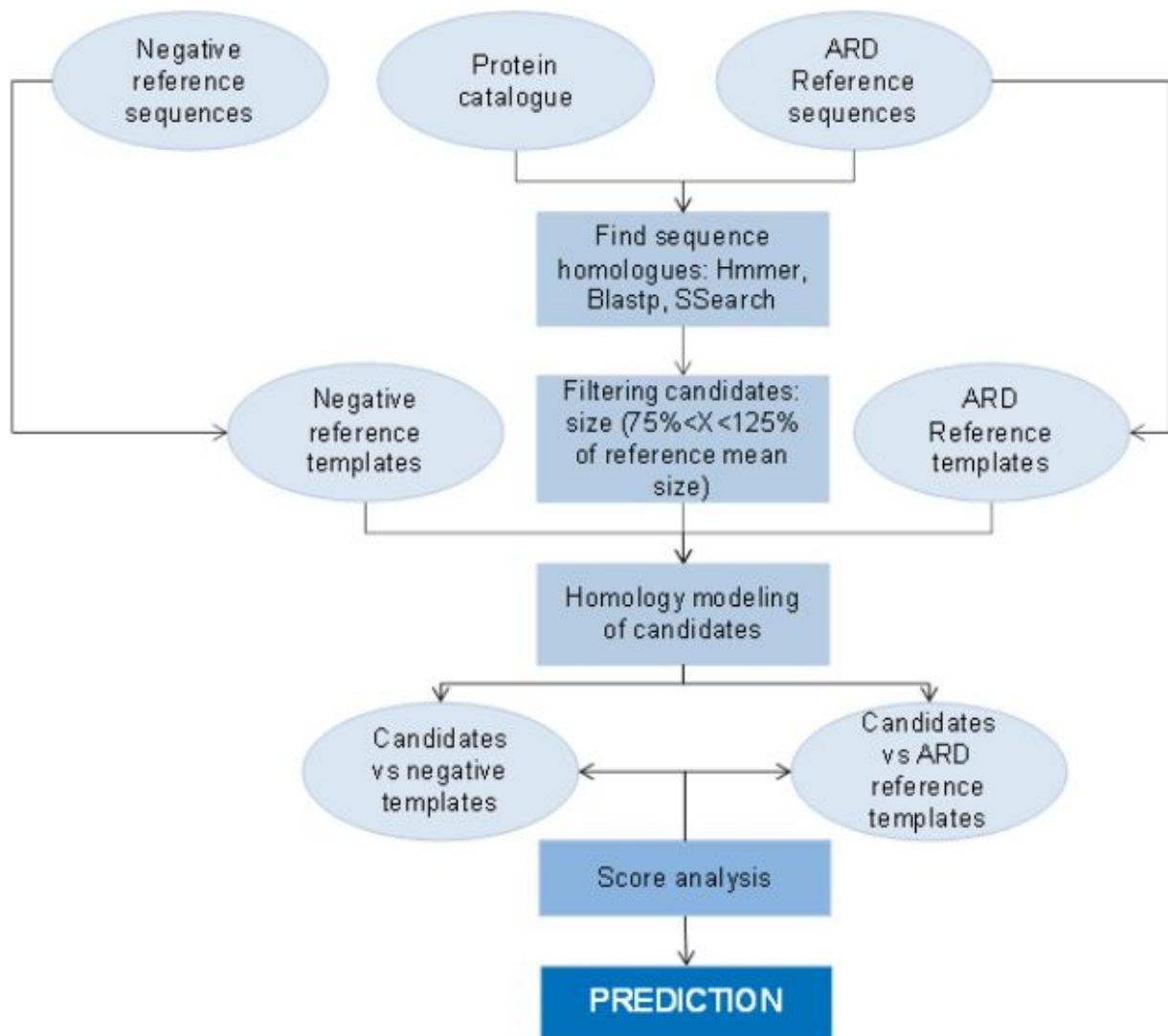
The example data appears in the history (2). We provided for the example: example_proteome (https://github.com/aghozlane/pcm/blob/master/example/example_proteome.faa).

This dataset correspond to the proteome of the following species:

Tree scale: 0.1



For this example we will search resistance genes in the proteome with the workflow of PCM:




Several parameters can be modified:

- One or several ARD families can be investigated. By default, PCM allows to screen the ARD families of clinical interest.
- Set the E-value threshold for candidate selection. This threshold impacts the candidates selection. A high e-value will select more candidates for PCM.
- Select the quality level of the modelisation. A high level of modelling quality will make PCM to spend more time to improve the protein model. By default, the mode fast is enough to produce a carbon-alpha model to predict an ARD.
- Select the number of model to calculate for modelling step. The homology modelling is a heuristic approach where the starting point is crucial. A high number of model corresponds to a high number of different start possible. By default, the number of model is set to 6 which is enough to screen the set of possibilities.
- Select the number of template to consider for modelling step. Setting this parameter higher makes the PCM more precise with a risk of overfitting. However some ARDs families do not have more than 3 ARD templates (notably MCR).
- Select the number of bootstrap to calculate for candidate classification. Setting this higher makes the classification more precise with an increase of calculation time.

Step by step:

1. Load the data:

Click on upload symbol  , then choose local files and finally start upload

Download from web or upload from disk





Regular

Composite

Collection


Rule-based

You added 1 file(s) to the queue. Add more files or click 'Start' to proceed.

Name	Size	Type		Genome	Settings	Status
 example_proteome.faa	14.7 MB	Auto-detect		unspecified (?)		0% 


Type (set all):


Auto-detect



Genome (set all):

unspecified (?)

 Choose local file

 Paste/Fetch data

Pause

Reset

Start

Close

2. Adjust your parameters set

Fasta file (protein sequence)



10: example_proteome.faa

Select an ARD family to investigate

☐ Select/Unselect all

- ☐ 16S rRNA methylase
- ☐ AAC(2')
- ☐ AAC(3)-I
- ☐ AAC(3)-II
- ☐ AAC(6')
- ☐ ANT
- ☐ APH
- ☐ Class A beta-lactamase
- ☐ Class B1 beta-lactamase
- ☐ Class B3 beta-lactamase
- ☐ Class C beta-lactamase
- ☐ Class D beta-lactamase
- ☐ DfrA
- ☐ Erm
- ☐ Fos
- ☐ Ldt
- ☐ MCR
- ☐ Qnr
- ☐ Sul
- ☐ Tet(M)
- ☐ Tet(X)
- ☐ Van ligase

E-value threshold for candidate selection

0.00001

--hfinder_evalue; Set the e-value threshold to search candidates; Setting this higher selects more candidates for the PCM default=1E-5

Quality level of the modelisation

Fast

--modelling_quality; Set the modelling quality level; Setting this higher makes the PCM more precise with an increase of calculation time; default=fast

Number of model to calculate for modelling step

6

--model; Setting this higher makes the PCM more precise with an increase of calculation time; default=6

Number of template to consider for modelling step

3

--template; Setting this higher makes the PCM more precise with a risk of overfitting; default=3

Number of bootstrap to calculate for candidate classification

10

--bootstrap; Setting this higher makes the PCM more precise with an increase of calculation time; default=3

Execute

3. Execute
4. Results are available when the right panel turn green. They can be visualised inside galaxy or downloaded (one by one).

**30: Output reference p
df output file PCM on d
ata 10**

**29: Output prediction
output file PCM on dat
a 10**

15 lines, 1 comments
format: **tsv**, database: ?

NEXTFLOW ~ version 19.01.0
Launching
'/pasteur/projets/policy01/galaxy-
prod/shed_tools/toolshed.pasteur.fr/
[angry_koch] - revision:
Saf9526318
[warm up] executor > slurm
[warm up] executor > local
Pro

1. Sequence	2. Type	3. Ref
Sequence	Type	Ref
CP002910_AEJ98074	Candidate_blaa	100
CP006919_AHM82220	Candidate_blaa	100
CP010881_AJO86874	Candidate_blaa	100
CP002910_AEJ99073	Candidate_blaa	0

**28: Output reference r
esult file PCM on data
10**

**27: Output pcm result
file PCM on data 10**

10: example_proteom

Sequence	Type	Ref	tneg	Adjusted p.value of being a Ref	TM.score_TMalign_ref	TM.score_TMalign_tneg	Prediction
Sequence	Type	Ref	tneg	Adjusted p.value of being a Ref	TM.score_TMalign_ref	TM.score_TMalign_tneg	Prediction
CP002910_AEJ98074	Candidate_blaa	100	0	3.46895869915534e-23	0.99744	0.81893	Very_likely_Ref
CP006919_AHM82220	Candidate_blaa	100	0	3.46895869915534e-23	0.98559	0.69116	Very_likely_Ref
CP010881_AJO86874	Candidate_blaa	100	0	3.46895869915534e-23	0.98686	0.86876	Very_likely_Ref
CP002910_AEJ99073	Candidate_blaa	0	100	1	0.91786	0.88424	Not_a_Ref
CP006918_AHM78534	Candidate_blaa	0	100	1	0.91917	0.88344	Not_a_Ref
CP010881_AJO86812	Candidate_blaa	100	0	3.46895869915534e-23	0.99862	0.88749	Very_likely_Ref
CP006918_AHM79651	Candidate_blaa	100	0	3.46895869915534e-23	0.99345	0.81105	Very_likely_Ref
AP012280_BAK88510	Candidate_blaa	0	100	1	0.88029	0.90219	Not_a_Ref
AE014075_AAN81121	Candidate_blaa	0	100	1	0.81273	0.87495	Not_a_Ref
CP010882_AJO86990	Candidate_blaa	100	0	3.46895869915534e-23	0.99862	0.88749	Very_likely_Ref
AE016830_AA081293	Candidate_blaa	100	0	3.46895869915534e-23	0.85035	0.61575	Likely_Ref
CP010876_AJO84151	Candidate_blaa	0	100	1	0.83077	0.88323	Not_a_Ref
CP010881_AJO86866	Candidate_blaa	100	0	3.46895869915534e-23	0.99831	0.86279	Very_likely_Ref
CP001918_ADF62979	Candidate_blaa	0	100	1	0.84524	0.86563	Not_a_Ref
CP010881_AJO86621	Candidate_blaa	100	0	3.46895869915534e-23	0.98686	0.86876	Very_likely_Ref

