Workflow IDseeqer

# Overview chart

Type of Data in the source database



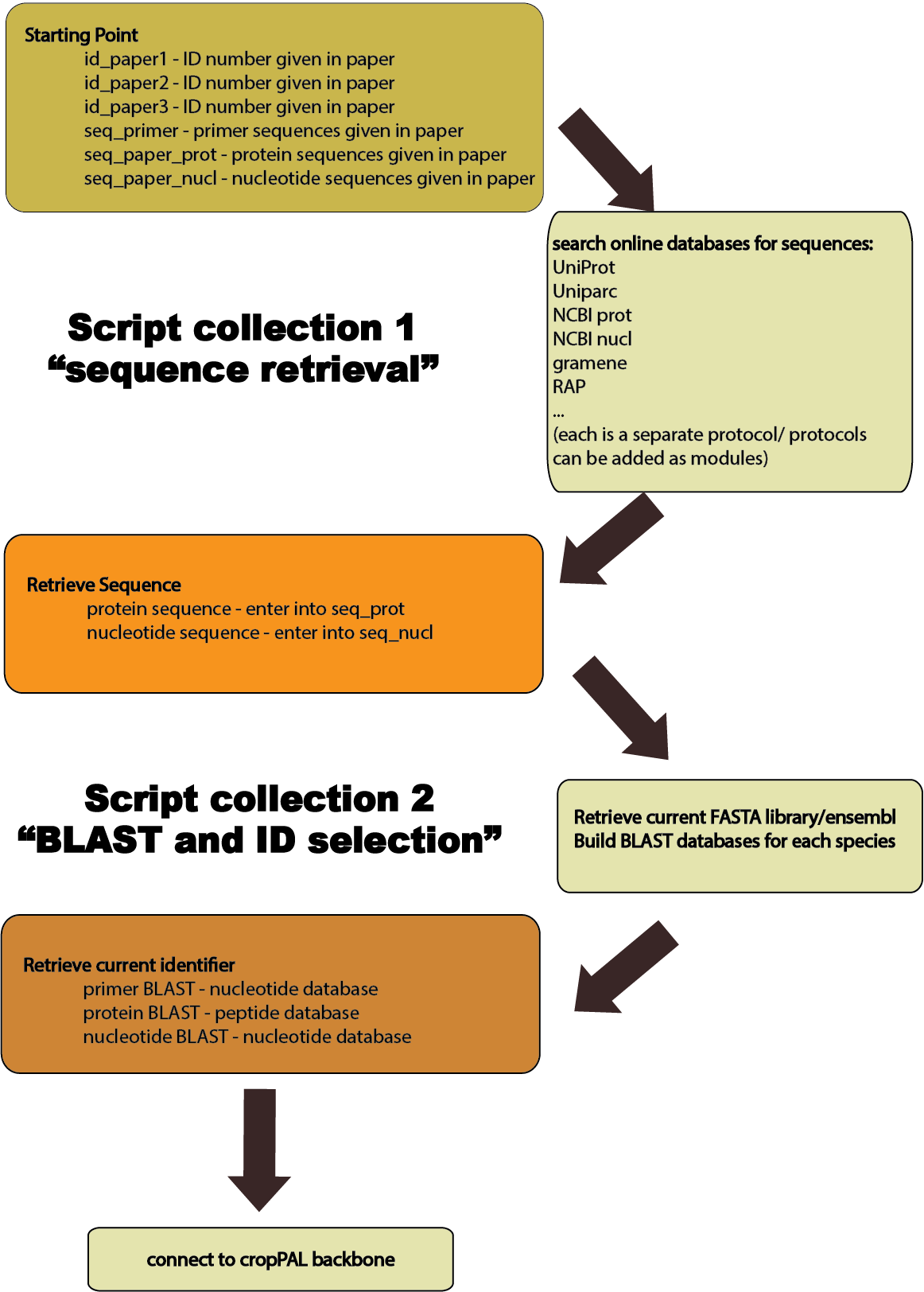
# Data filling

The curation data file must contain enough information to search for a sequence or provide a sequence in order to connect the data point.

If no sequence data is given the ids collected from the curation will be sent off to databases to retrieve sequences.

At this point several sequences can be retrieved. That means each data point can have a primer sequence, protein sequence, nucleotide sequence as well as ids. In the end the locus from difference sources can be compared if we want.

# Pipeline flowchart



# Sequence Retrieval Hierarchy

The program takes on entry (row ID) and collects id\_paper1

The id\_paper1 will be sent to the database in the following order:

Uniport

NCBI prot

Gramene

RAP

NCBI nucl

NUCCORE

Uniparc

If at any of these stages a sequence is retrieved and the database is updated: Stop the search

# BLAST retrieval Hierarchy

Once all the sequences are retrieved there is another program that will start the ID retrieval (BLAST)

The program will obtain the sequence X and use the program Y in this order:

|  |  |  |  |
| --- | --- | --- | --- |
| order | Type of sequence (X) | Program (Y) | update column for final output |
| 1 | seq\_primer | Blast\_primer | Locus\_primer |
| 2 | Seq\_paper\_prot | BLAST\_prot | Locus\_seq\_paper |
| 3 | Seq\_paper\_nucl | BLAST\_nucl | Locus\_seq\_paper |
| 4 | Seq\_prot | BLAST\_prot | Locus\_seq |
| 5 | Seq\_nucl | BLAST\_nucl | Locus\_seq |