▷ Adjust argparse such that boolean input no longer needs the ~~keyword~~ value "True":

"--taxonomycheck True" ⇨ "--taxonomycheck"

▷ Implement ~~reference~~ integration of References.

■ Make modific.:
For each qualifier you have, make a column

5' CDS location needs a "plus 1".
5' CDS incorrect 5' CDS correct
667 ⇒ 668

* AT LEAST FOR CHECKLISTS *

2.) Remove "transl_table" from source features; if anything, it should be in a CDS feature.

3.) Make sure that fourth field of title line ~~is equal to content~~ has same content as as soure feature "mol_type".

Falscher Output bei Col. "INTRON"

Checklists for mafk in Pyrus:

- fmk_intron has erroneous name ← id = "K" (instead of id = "fmk")
- Intron wird nicht erkannt!

Before and after Positions are specified where there should be Exact Positions
↳ Falscher Output bei Col. "5'_PARTIAL" "3'_PARTIAL"

Why is there a join(1..41, 43..530)?

It appears that the misc-feature length calculation is wrong.

6.5.

6.5.1. ! (Location object!)

6.5.3.

---

make decision of location-change based on seq-loc-comparison (Before/after removal of ambiguities)

If leading and trailing ~~space~~ chars indeed removed, feature location must be adjusted

1..501 => >1..<501

! [6.] Remove data that have only missing data.

7.) Set up list of source qualifiers that can be accessed from anywhere in the software.

1.) Remove all missing segs per alignment

1.) Check that all taxon_names in .nex also in .csv

ANNONEX2 EMBL → Rename to: "nex2ena-float" / "nex2ena-check"

5.) ~~Add~~ Specify "data class" (3.1. in ENA manual) (STD)

TO DO - LONG TERM

1.) Rename "Degapping Ops" with "Cleanup Ops"

- I changed Taxon_3 in TestData_1.nex by making the non-coding ends "?". This must be replaced in the corresponding figure.

- ~~Update the figures~~

- The ID line changed in all examples (TestData_1, TestData_2). Check if this affects the figures based on these datasets in any way.

In Manuskript:

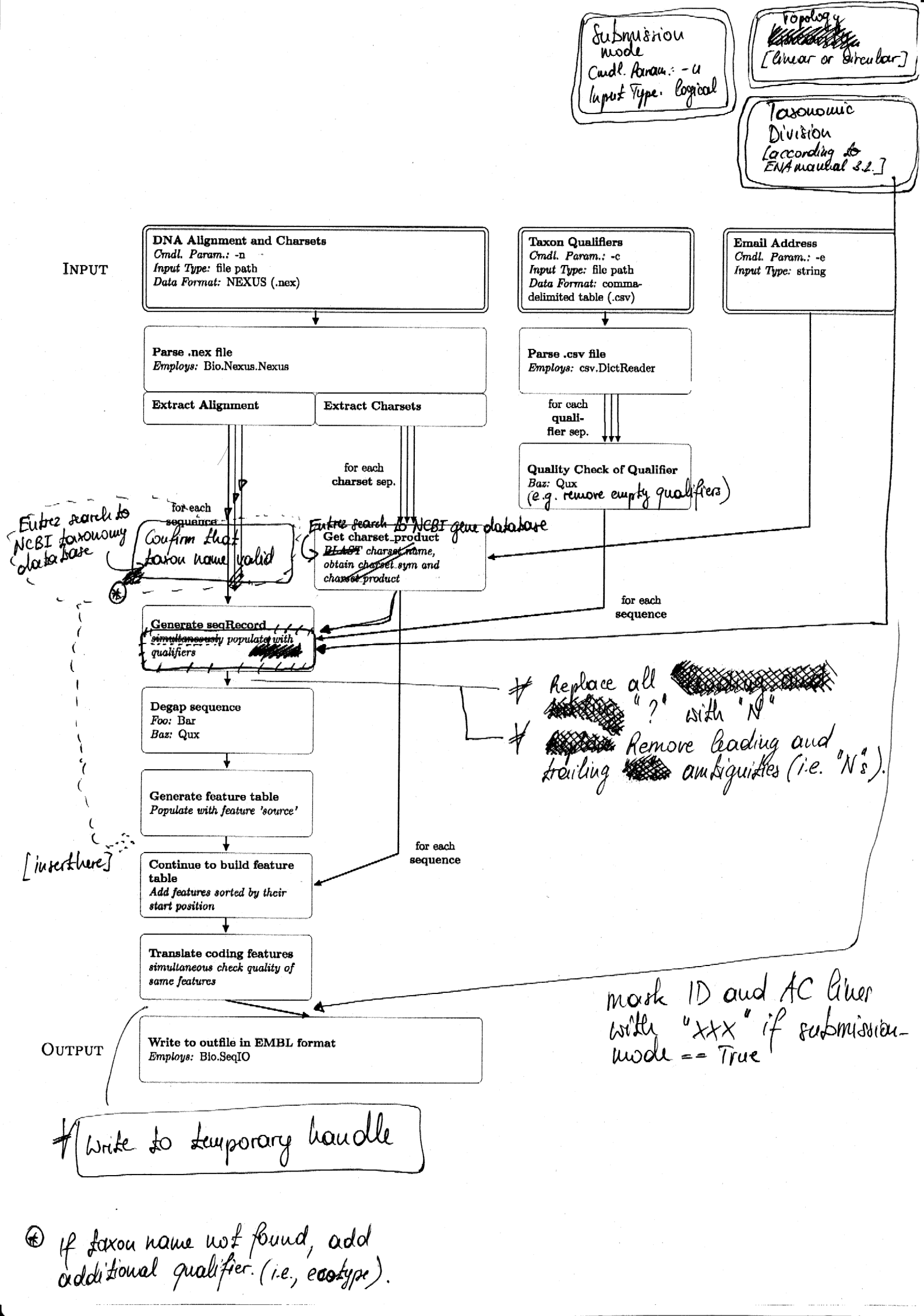
- ☐ Write that charset names must have a special format:

charsetName _ charsetType

(_ = underscore)

charsetName → BLASTED FOR charset-cleaning

charsetType → must be part of ISIC (RNA, gene, intron, ...)


Submission mode
Cmdl. Param.: -u
Input Type: logical
Topology
[Linear or Circular]
Taxonomic Division
[according to EMBL manual 3.2.]
INPUT
DNA Alignment and Charsets
Cmdl. Param.: -n
Input Type: file path
Data Format: NEXUS (.nex)
Taxon Qualifiers
Cmdl. Param.: -c
Input Type: file path
Data Format: comma-delimited table (.csv)
Email Address
Cmdl. Param.: -e
Input Type: string
Parse .nex file
Employs: Bio.Nexus.Nexus
Parse .csv file
Employs: csv.DictReader
Extract Alignment
Extract Charsets
for each qualifier sep.
for each charset sep.
Quality Check of Qualifier
Baz: Qux
(e.g. remove empty qualifiers)
for each sequence
Entrez search to NCBI taxonomy database
Confirm that taxon name valid
Entrez search to NCBI gene database
Get charset_product
BLAST charset name, obtain charset sym and charset product
Generate seqRecord
simultaneously populate with qualifiers
for each sequence
Degap sequence
Foo: Bar
Baz: Qux
Replace all "?" with "N"
Remove leading and trailing ambiguities (i.e. "N"s).
Generate feature table
Populate with feature 'source'
[insert here]
Continue to build feature table
Add features sorted by their start position
for each sequence
Translate coding features
simultaneous check quality of same features
mask ID and AC lines with "XXX" if submission-mode == True
OUTPUT
Write to outfile in EMBL format
Employs: Bio.SeqIO
Write to temporary handle
* If taxon name not found, add additional qualifier. (i.e., ecotype).