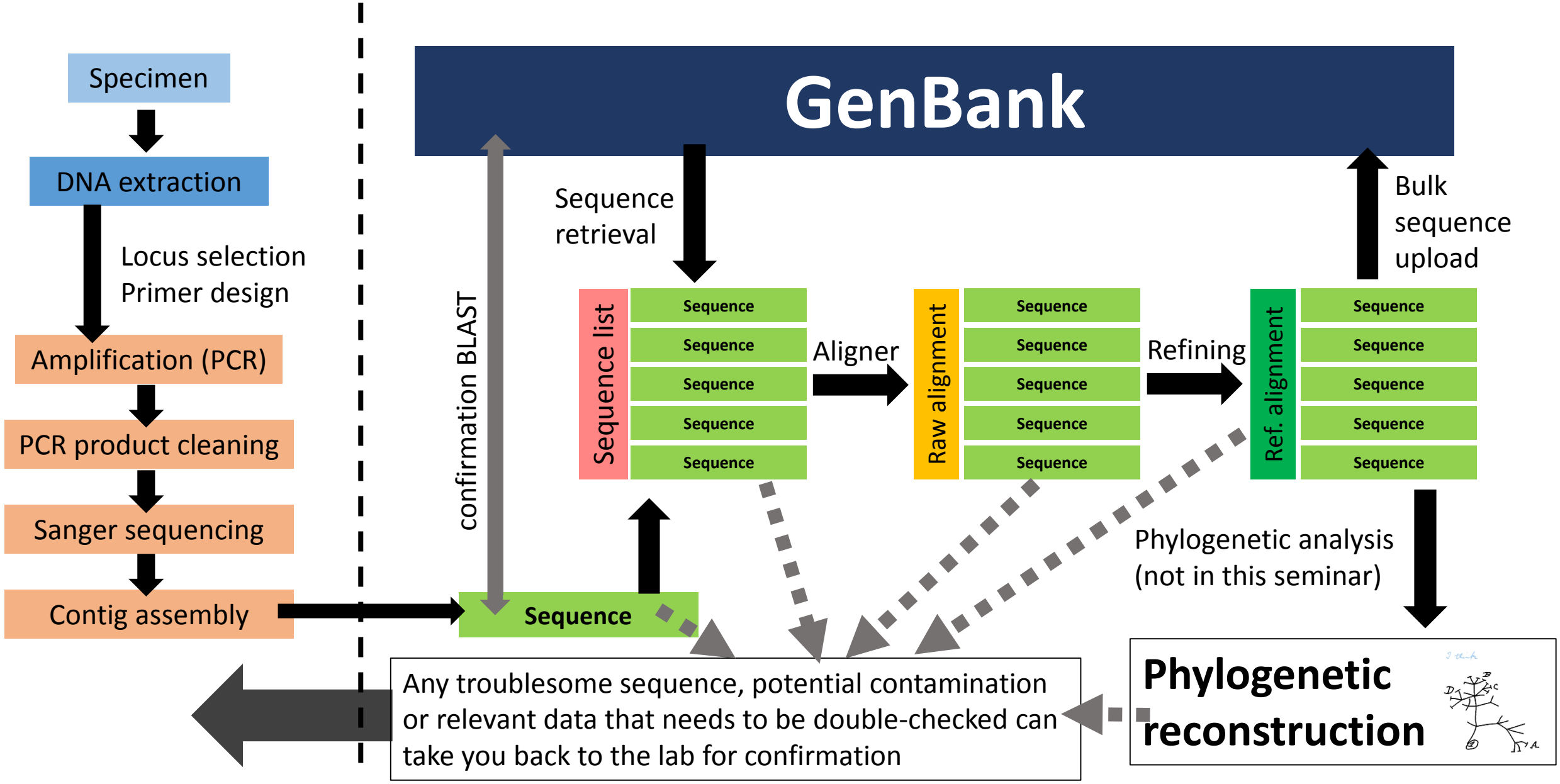
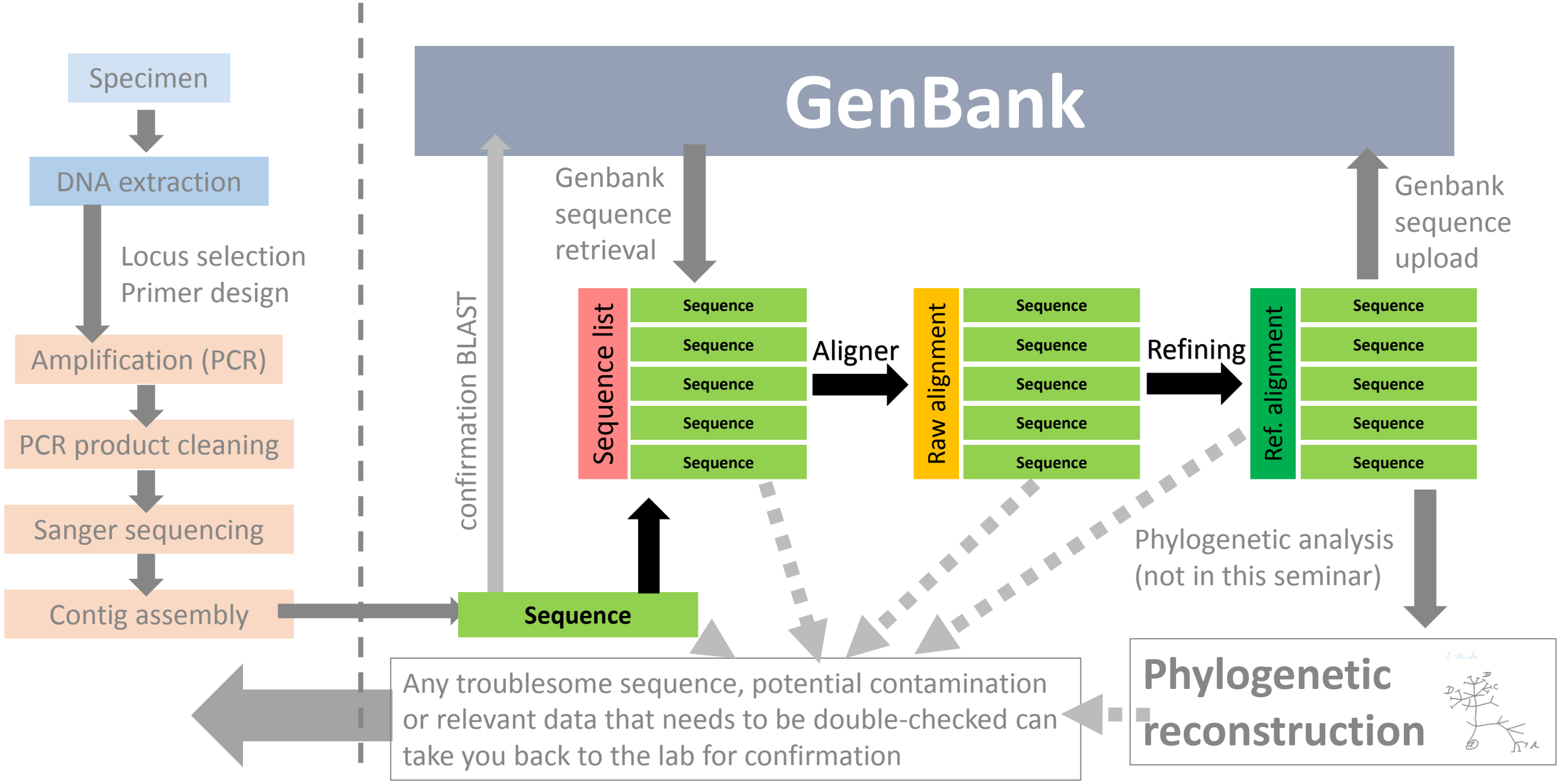
Sequence formats. Alignments

Rafael Medina (rafael.medina.bry@gmail.com)
Yang Liu (yang.liu@uconn.edu)

A road map to phylogenetic data gathering



Today: alignments



Composing a sequence list

Never use a Word-type text editor for sequence manipulation, choose instead plain and versatile editors such as those used in programming with constant character width

Our choices:

Notepad++ (Win)

Wrangler (Mac)

Vim (Linux)

Sequence in fasta format (extension .fas or .fasta)

Name of the sequence after the “>”

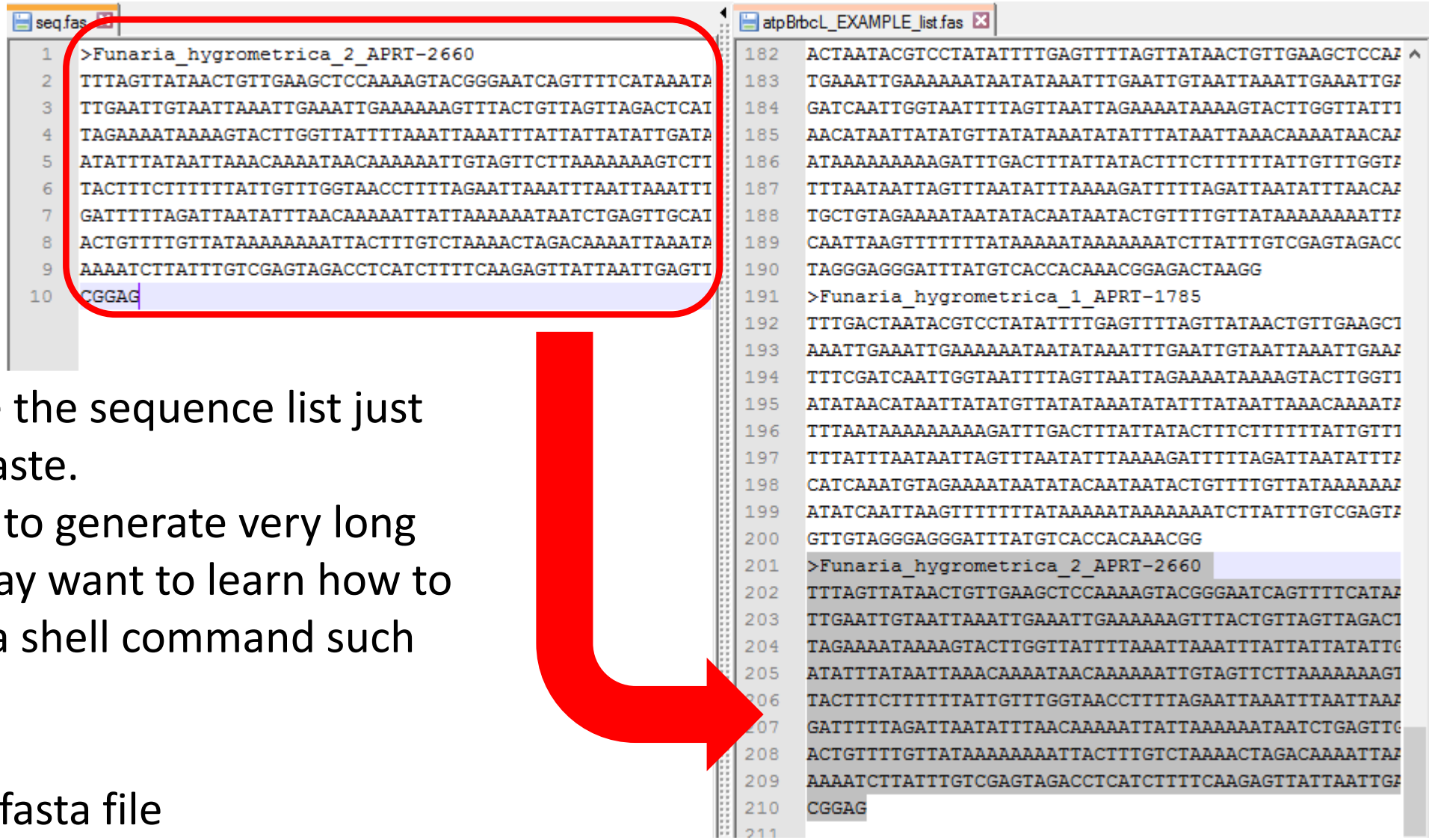
seq.fas

```
1  >Funaria_hygrometrica_2_APRT-2660
2  TTTAGTTATAACTGTTGAAGCTCCAAAAGTACGGGAATCAGTTTTCATAAATA
3  TTGAATTGTAATTAAATTGAAATTGAAAAAAGTTTACTGTTAGTTAGACTCAT
4  TAGAAAATAAAAGTACTTGGTTATTTTAAATTAAATTTATTATTATATTGATA
5  ATATTTATAATTAAACAAAATAACAAAAAATTGTAGTTCTTAAAAAAAGTCTT
6  TACTTTCTTTTTTATTGTTTGGTAACCTTTTAGAATTAAATTTAATTAAATTT
7  GATTTTTAGATTAATATTTAACAAAAATTATTAAAAAATAATCTGAGTTGCAT
8  ACTGTTTTGTTATAAAAAAAATTACTTTGTCTAAAACTAGACAAAATTAAATA
9  AAAATCTTATTTGTCGAGTAGACCTCATCTTTTCAAGAGTTATTAATTGAGTT
10 CGGAG
```

The nucleotide sequence body comes next

There are several sequence and alignment formats (more on this later)

Composing a sequence list

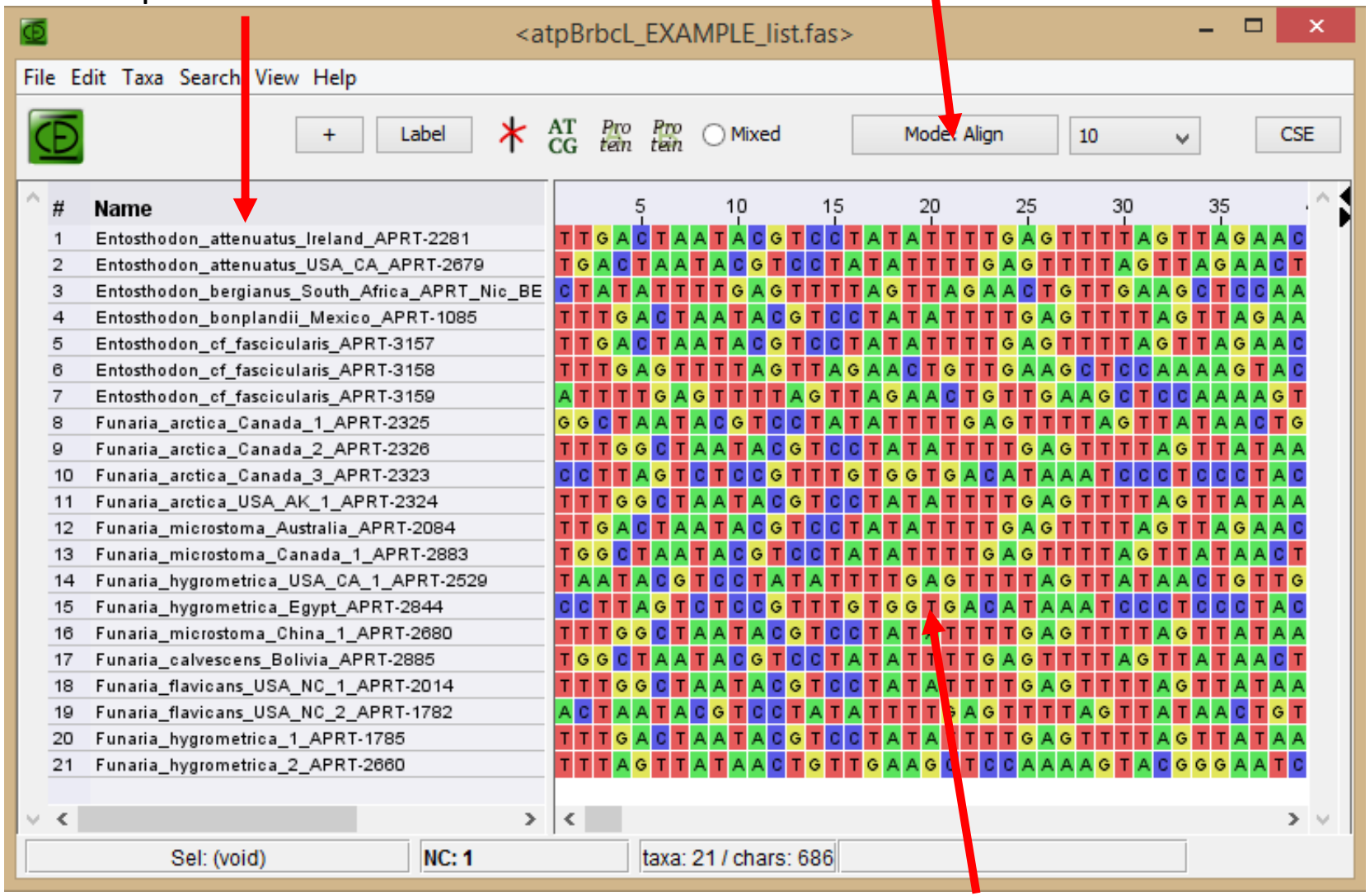


To generate the sequence list just copy and paste.
If you need to generate very long lists, you may want to learn how to do it using a shell command such as `cat`

Save it as a fasta file

Alignment editor: PhyDE

- Intuitive, versatile and visual alignment editor
- Runs on java (system independent)
- Has several interesting plugins
- It's free!



Aligner: MUSCLE

Although you can align manually easy sequence lists, it will save you time to use some aligner like MUSCLE

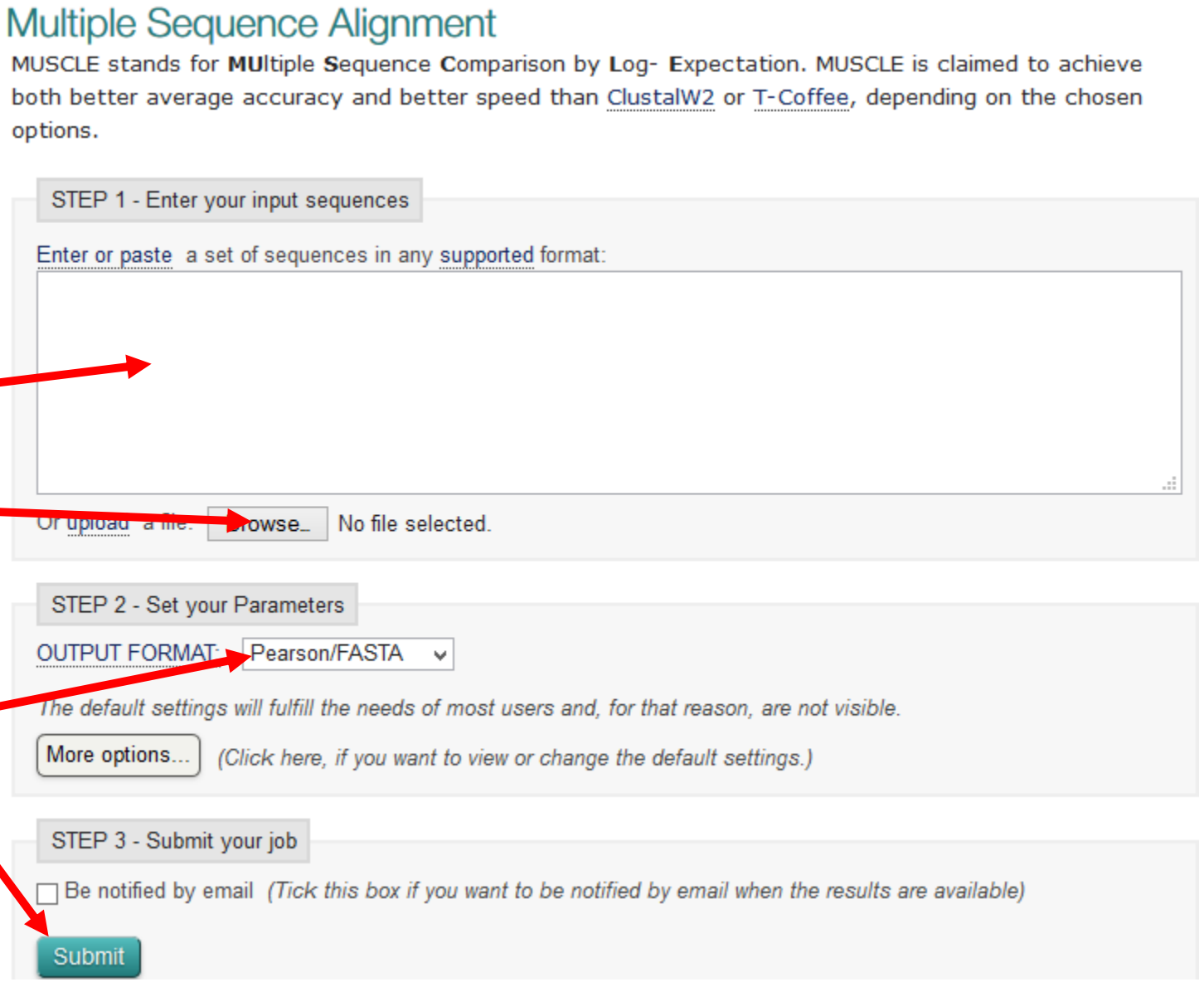
Alternatives:

- Clustal
- Mafft
- Tcoffee

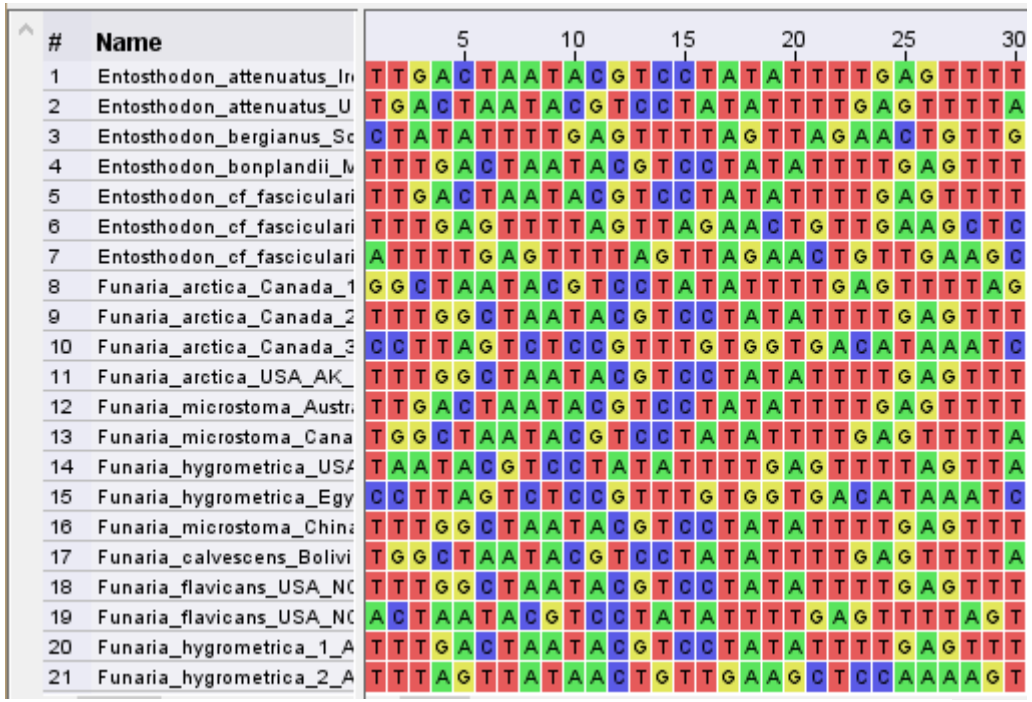
paste or upload your fasta list

Fasta format

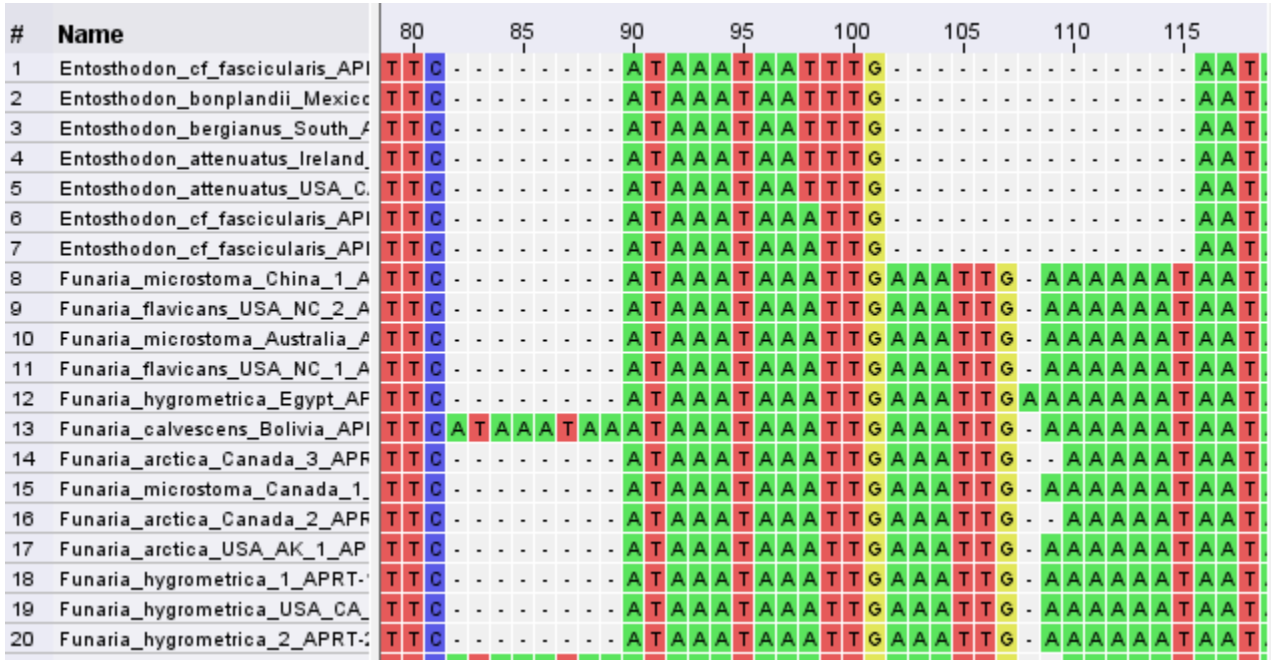
Then, run the aligner



Sequence list vs. Alignment: an important reminder

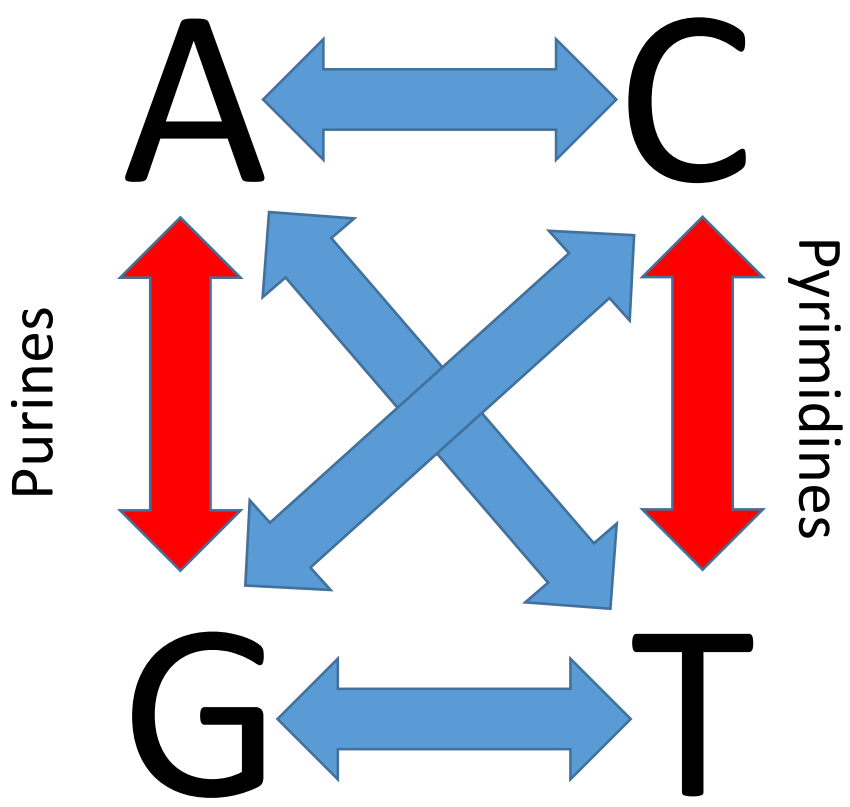


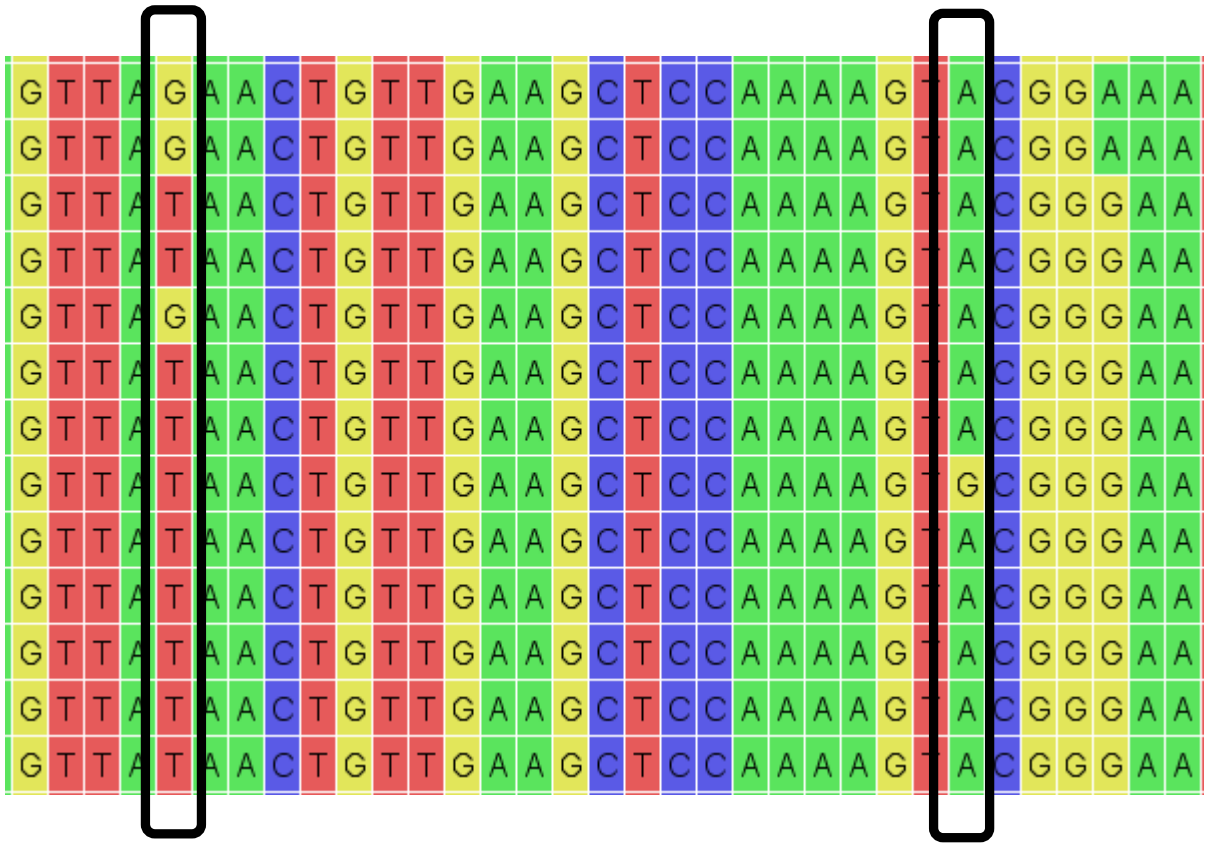
(Assuming there were no sequencing errors) these are plain, objective data



However, an alignment is more tan plain data, it is also a statement on nucleotide homology: you are assuming (either if you are right or not) that each position (column) contains homologous nucleotides, and this will become VERY relevant in phylogenetic analysis. Be aware of it! This is why the manual adjustment or refining of the alignment is important

Alignment elements: substitutions





Transversion

Transition

Alignment elements: gaps

Gaps caused by INsertions or DELetions (indels)

You can easily add an indel block to you alignment using the software SeqState

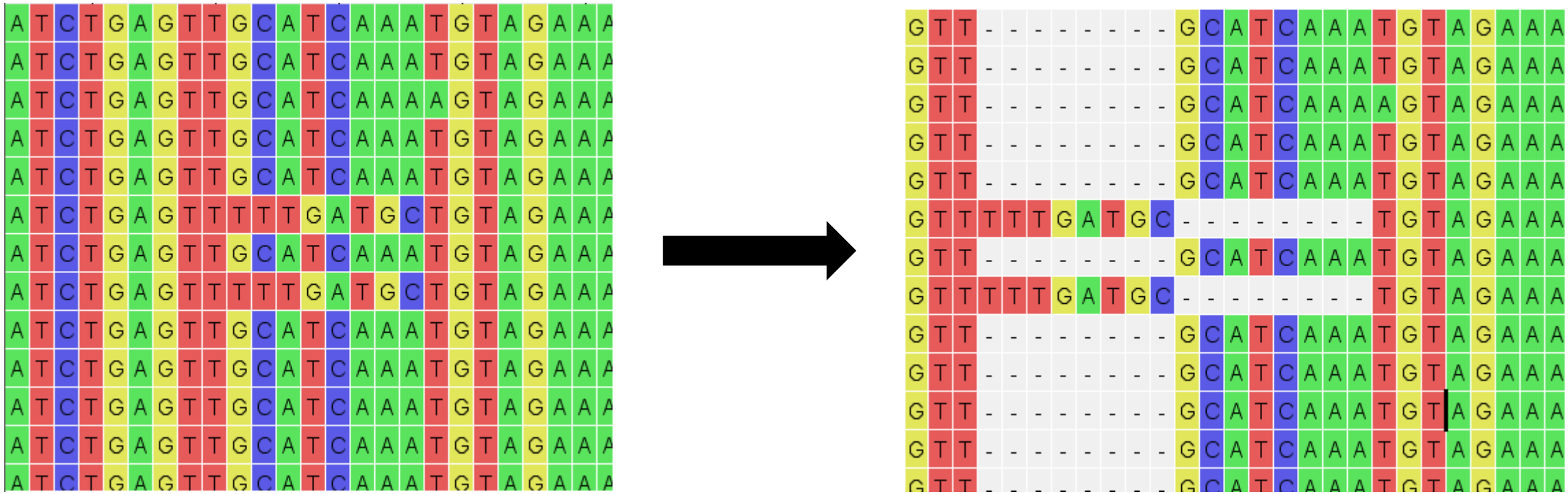
Read about different ways to code indels in: Simmons & Ochoterena 2000 Gaps as Characters in Sequence-Based Phylogenetic Analyses. *Syst. Bio*. 49: 369-381

Gaps are often ignored in phylogenetic analyses, but they can be evolutionary informative

If you want to acknowledge this information, you can code indels as binary characters in an adjacent block

Alignment elements: inversions

This is actually an inversion, not 8 substitutions



Exclude or code these features, but do not keep them in their raw state or they will overestimate the divergence of these two sequences

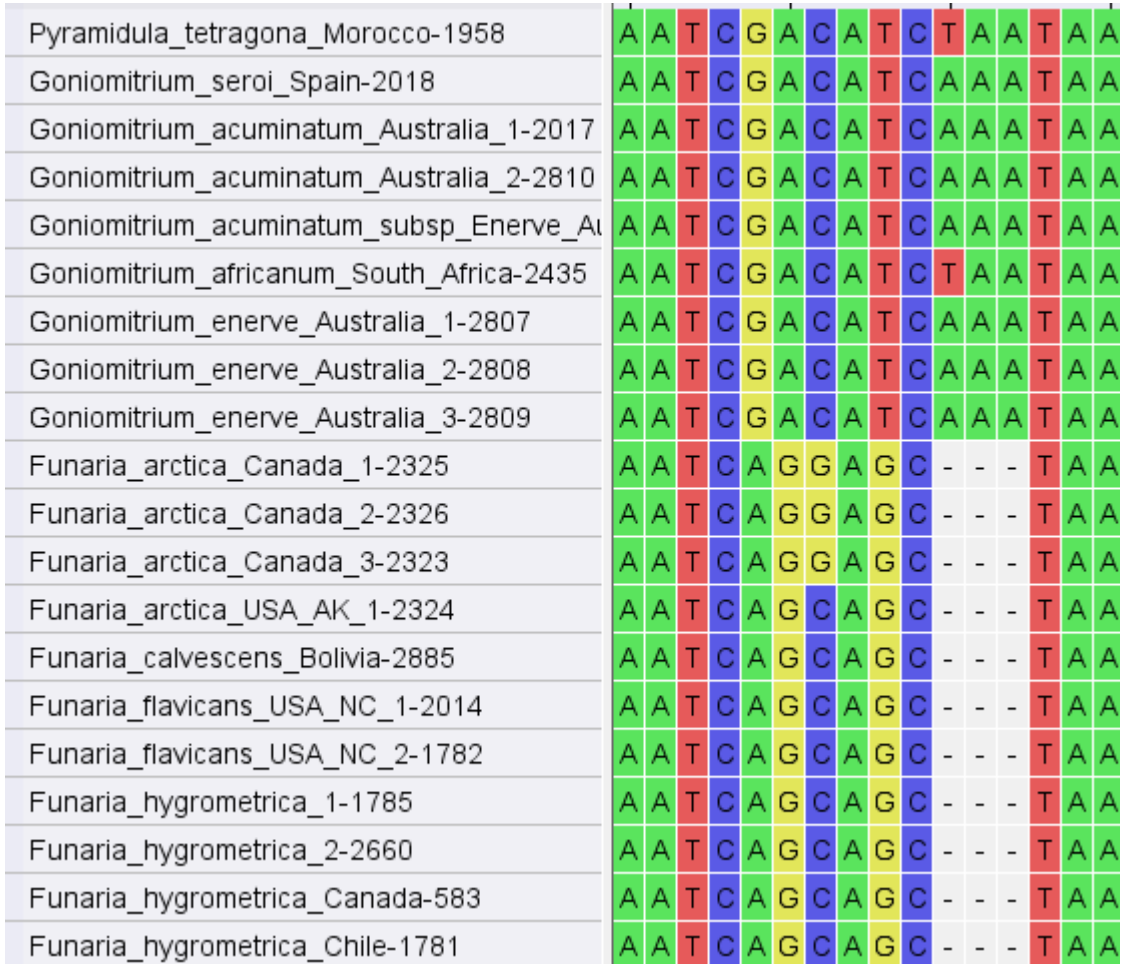
Alignment elements: repeated elements

Some repeated features, such as microsatellites, may be of interest for phylogenetic analysis too, but confirm them, if necessary, with your raw data

Alignment elements: indels in coding regions

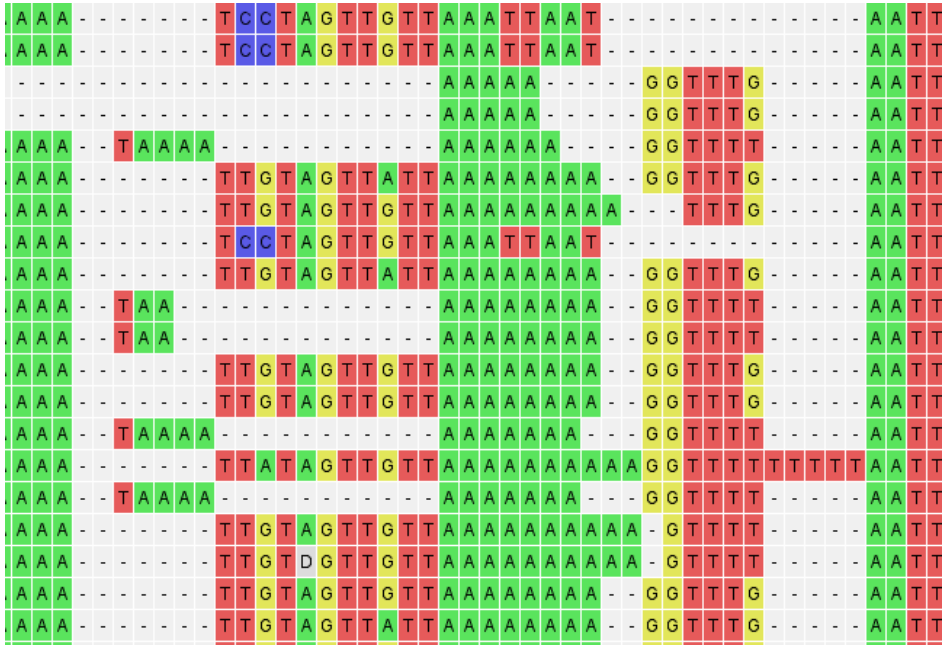
Be always suspicious about shifts in reading frames or very divergent sequences. You may need to check again your raw chromatograms or even re-sequence that sample

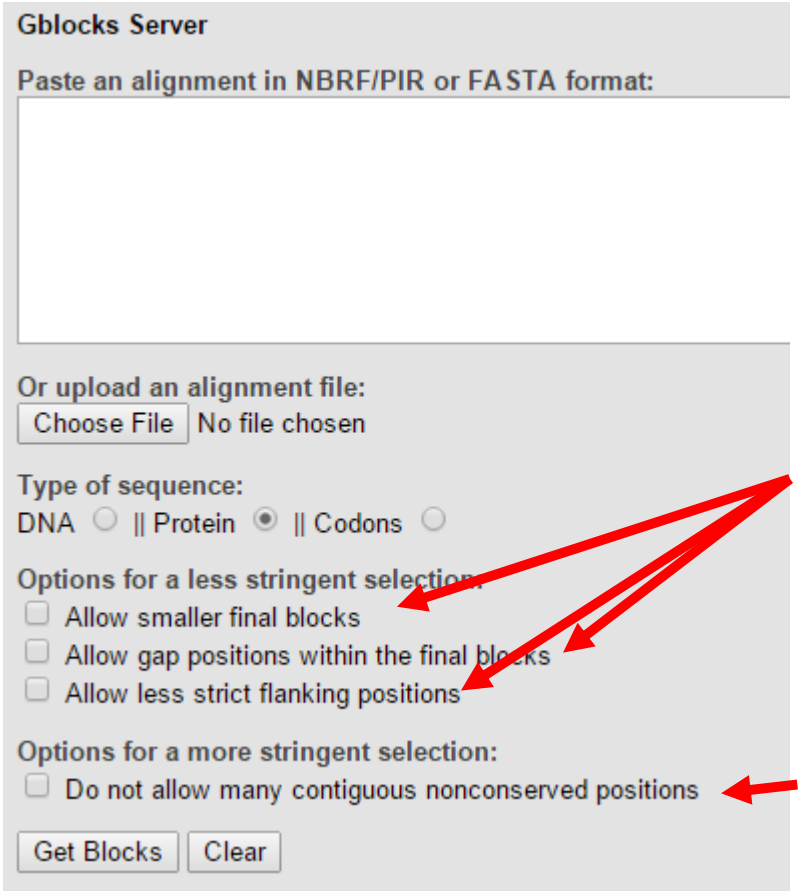
However, 3x nucleotide indels in coding regions are indeed a possibility (they would respect the reading frame)



Alignment elements: ambiguity

You cannot get rid of the ambiguity in many alignments. The best you can do is to exclude the areas of the alignment that are not trustworthy, either manually or using softwares such as Gblocks
It is always better to ignore dubious data than to introduce noise in your analysis. If you are not sure about the homology: exclude





Less stringent options allow more flexibility

You can increase the stringency to get a more conservative alignment

"Exclude" means that an area of the alignment will not be analyzed. Never erase manually any nucleotide, unless you confirm it with the raw chromatogram

Alignment elements: ambiguity

```
Flank positions of the 10 selected block(s)
Flanks: [10  81]  [90  107]  [109  193]  [195  248]  [250  341]  [356  465]  [470  481]  [489  546]  [555  719]  [721  727]
```

List of blocks to be kept

Visualization

Parameter summary

```
Parameters used
Minimum Number Of Sequences For A Conserved Position: 11
Minimum Number Of Sequences For A Flanking Position: 11
Maximum Number Of Contiguous Nonconserved Positions: 8
Minimum Length Of A Block: 5
Allowed Gap Positions: With Half
```

You can download the corrected alignment without the exclusion blocks, however it is better to keep the sequences unedited. You will be able to exclude parts of the alignment before the analysis, just keep a record of the exclusion blocks.

Sequence and alignment formats

FASTA format

Extension: .fas or .fasta

Each sequence starts with “>”

Poor in information

Sequence name

Sequence and alignment formats

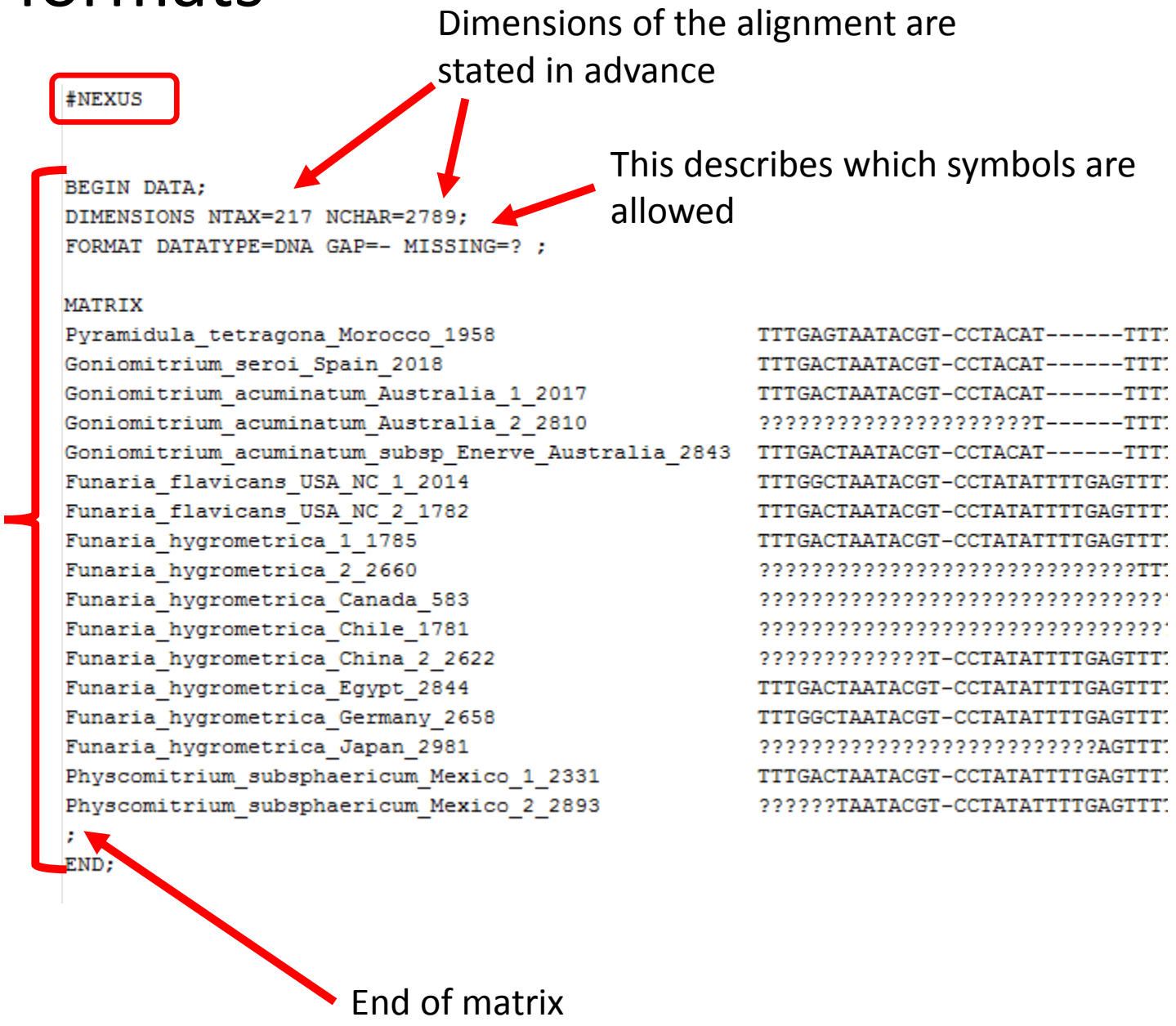
NEXUS format

Extension: .nex

The whole file starts with #NEXUS

They are structured in blocks of information, each of them starting with “BEGIN” and ending with “END;” (Don’t forget the “;”s)

Alignments are usual referred as “MATRIX”. You can encode indels, exclude areas, include specific information for programs, etc



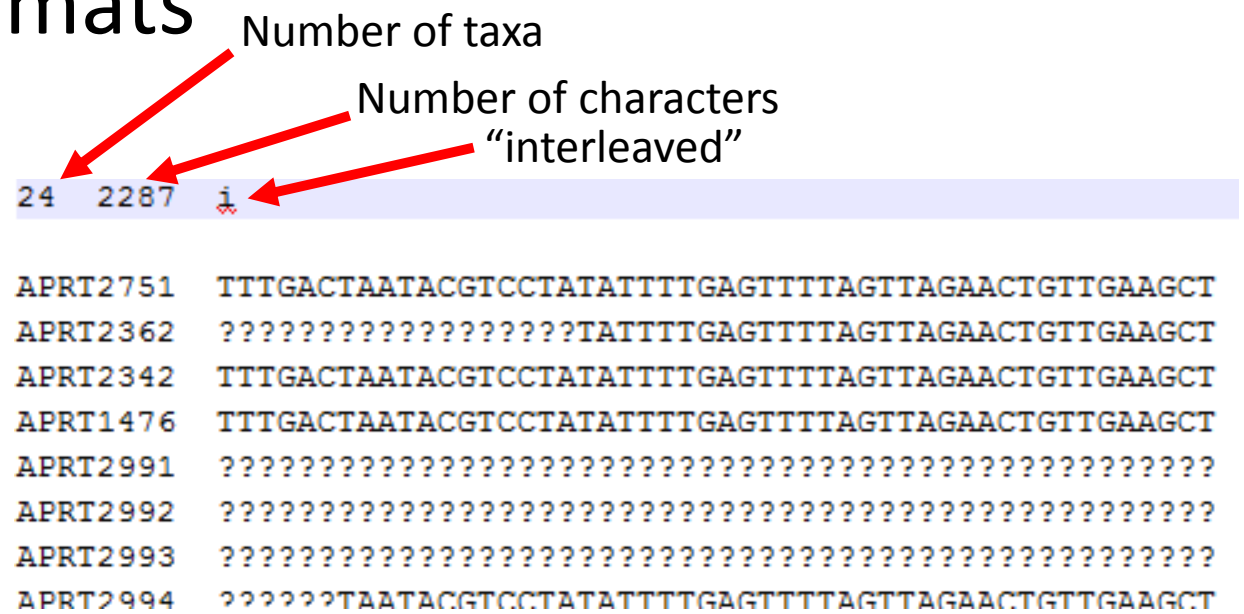

```
#NEXUS

BEGIN DATA;
DIMENSIONS NTAX=217 NCHAR=2789;
FORMAT DATATYPE=DNA GAP=- MISSING=? ;

MATRIX
Pyramidula_tetragona_Morocco_1958                    TTTGAGTAATACGT-CCTACAT------TTT
Goniomitrium_seroi_Spain_2018                        TTTGACTAATACGT-CCTACAT------TTT
Goniomitrium_acuminatum_Australia_1_2017             TTTGACTAATACGT-CCTACAT------TTT
Goniomitrium_acuminatum_Australia_2_2810             ??????????????????????T------TTT
Goniomitrium_acuminatum_subsp_Enerve_Australia_2843  TTTGACTAATACGT-CCTACAT------TTT
Funaria_flavicans_USA_NC_1_2014                      TTTGGCTAATACGT-CCTATATTTTGAGTTT
Funaria_flavicans_USA_NC_2_1782                      TTTGACTAATACGT-CCTATATTTTGAGTTT
Funaria_hygrometrica_1_1785                          TTTGACTAATACGT-CCTATATTTTGAGTTT
Funaria_hygrometrica_2_2660                          ??????????????????????????????TT
Funaria_hygrometrica_Canada_583                      ????????????????????????????????
Funaria_hygrometrica_Chile_1781                      ????????????????????????????????
Funaria_hygrometrica_China_2_2622                    ????????????????T-CCTATATTTTGAGTTT
Funaria_hygrometrica_Egypt_2844                      TTTGACTAATACGT-CCTATATTTTGAGTTT
Funaria_hygrometrica_Germany_2658                    TTTGGCTAATACGT-CCTATATTTTGAGTTT
Funaria_hygrometrica_Japan_2981                      ??????????????????????????AGTTT
Physcomitrium_subsphaericum_Mexico_1_2331            TTTGACTAATACGT-CCTATATTTTGAGTTT
Physcomitrium_subsphaericum_Mexico_2_2893            ??????TAATACGT-CCTATATTTTGAGTTT
;
END;
```

Sequence and alignment formats

PHYLIP format

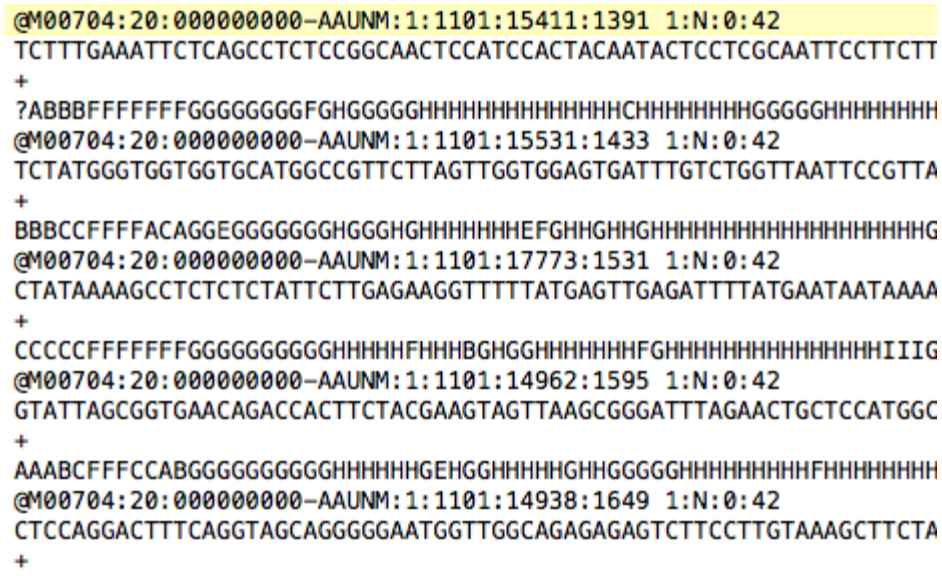
Extensions: .phy .phylipi

This format usually has a lot of restrictions (like a limit in the sequence name length), but some softwares only read PHYLIP



FASTAQ format

Complex, data-rich format specific for High-Throughput Sequencing. It contains quality scores for each base



Format conversion

- PhyDE allows to export an alignment in fasta and nexus format
- Geneious provides a wider spectrum of formats and it is a good resource if you need a phylip file
- Mesquite is also a free, java-based program that you may use for a lot of purposes, including format conversion
- Alternatively, you can use the format converter of HIV database