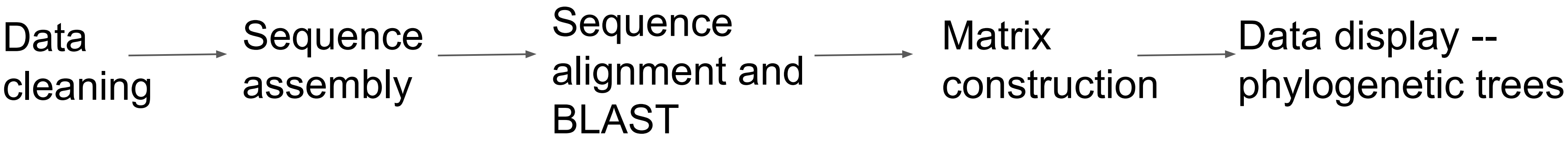
Reviewing our pipeline...



Data cleaning

Sequence assembly

Sequence alignment and BLAST

Matrix construction

Data display -- phylogenetic trees

Data cleaning

- Pre-processing now saves tons of time later! What were some common errors/noise that we tried to account for?
    - Missing / gapped data
    - Frameshift / changing reading frames
    - Premature stop codons
    - Making sure we’re using the right translation table! (What’s the correct translation table for corals/anemones?)

Sequence assembly

- Remember the assembly scripts you wrote -- what were some different strategies for approaching the assembly problem?
- What was the ‘trick’ that we eventually used to write our script? (Hint: after preprocessing with the .find() function, we were able to use this trick!)
- What was the big-O cost of this function? (Hint: we exhaustively searched every possible assembly…)
- What are deBruijn graphs? What was the solution to the Bridges of Konigsberg problem?

Sequence alignment and BLAST

- What does BLAST do? Is it the most accurate algorithm of its kind? (Hint: think about tradeoffs between speed and accuracy)
- Think about how you’d go about BLASTing a sequence through Biopython…
- What are e-values?
- What are some different ways to compare sequence alignments? (Hint: ‘turning a hen into a fox’!)

Matrix construction

- What are different ways to score similarity/differences between taxa?
- What metric did we use for our own distance matrix function?

Data display -- phylogenetic trees

- What are some different methods for tree construction?
- What do you recall from…
  - That Khan academy video?
  - Our candy phylogeny activity?
  - Tree-building and tree-type research last Thursday?

What’s next?

You’ll be applying what you’ve learned from the past three months to a large, unwieldy data set of sequence data from deep-sea organisms!

Each group will receive a different set of three taxa (organisms); you’ll receive three genes from each taxa.

In your pairs, you’ll clean and analyze your taxa. As a group, we’ll build distance matrices and trees together!