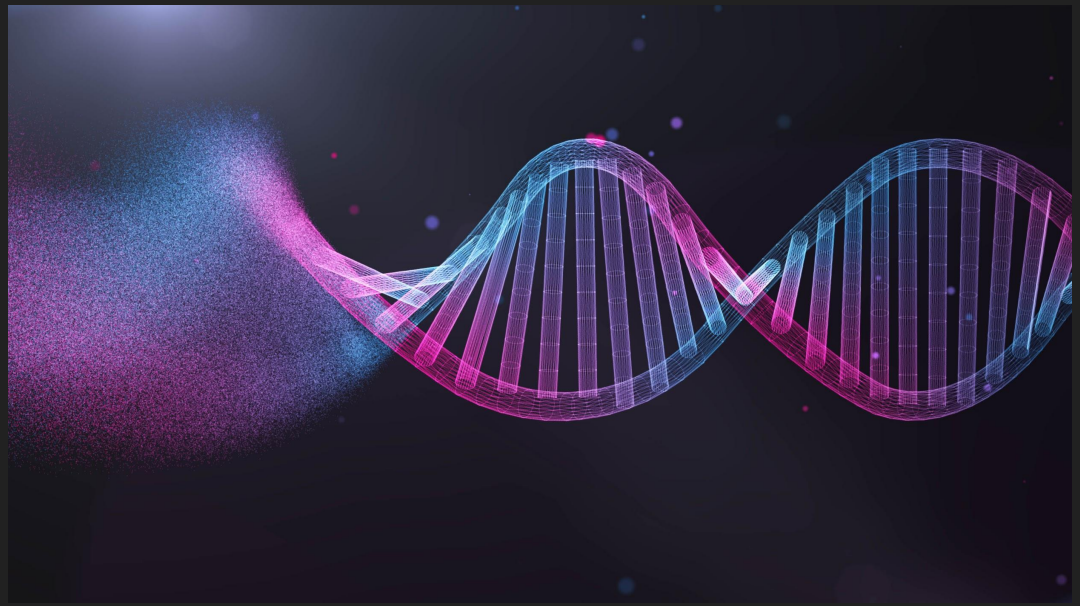
Computer-assisted sequence analysis



What is computer assisted sequence analysis?

Data stored in protein databases (NCBI, Swiss-Prot, ...etc) and nucleic acid databases (EMBL, GenBank, ...etc).

Algorithms for comparing similar amino acid and nucleic acid sequences (Needleman-Wunsch, Smith-Waterman, ...etc).

Tools and Programs for ease of use (Benchling, Genome Workbench, ClustalW, BLAST, ...etc).

Benchling

CLUSTAL

What does this do?

This allows for efficient comparison of sequences of amino acids, DNA, and RNA.

The original algorithms were described by Saul Needleman and Christian Wunsch in 1969.

These algorithms simplify the calculations of similarity between sequences and simplify the alignment of these sequences based on their similarities.

Algorithm basics:

Quality of alignment is determined by alignment score.

Alignment Score is determined by cost-benefit sum analysis based on:

1. The number of similarly aligned residues.
2. The number of differently aligned residues.
3. The number of gaps created to align sequences.

1. ATGAA
   ⋮ ⋮ ⋮
   ATG T

   There are 3 pairs of correctly aligned residues.

2. ATGAA
   ⋮
   ATG T

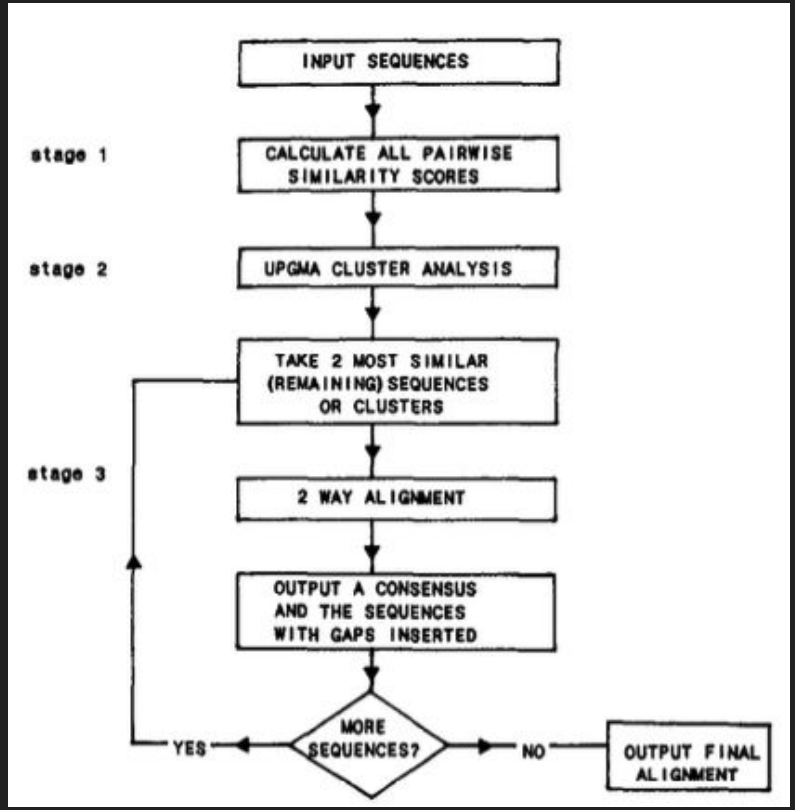
   There is 1 pair of incorrectly aligned residues.

3. ATGAA
   ⋮
   ATG T

   1 gap was created to align the sequences.

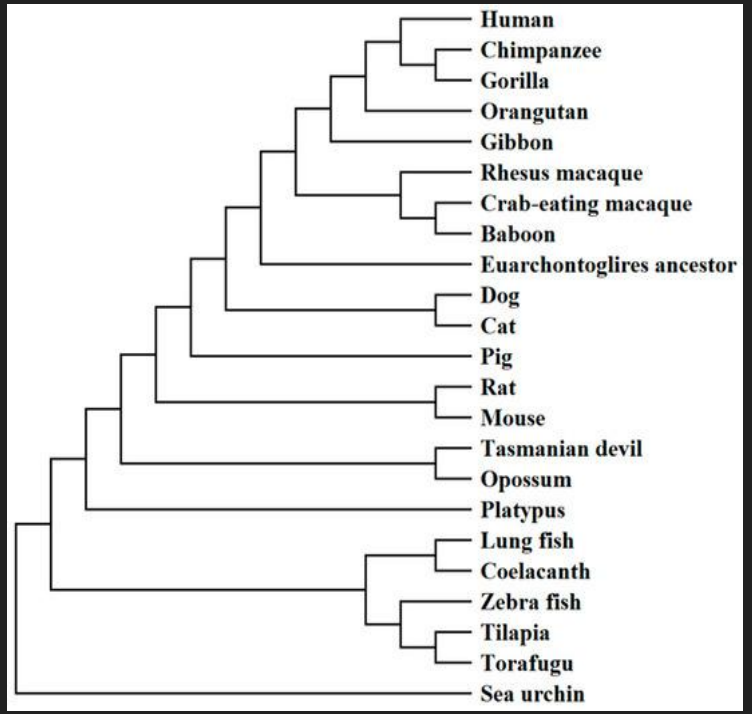
Algorithm in practice

1. Every sequence in a set of N sequences is compared to every other sequence. This is calculated through sets consecutively aligned residues and the number of gaps in the alignment.
2. The similarity between sequences is mapped into a phylogenetic tree based on similarity.
3. Sequences are aligned based on the phylogenetic tree, and sconservative residues are taken into account.



How are multiple sequence alignments used in other applications?

- Drug design
- Characterization of genetic diseases
- Phylogenetic relationships
- Historical sequence profiling



When did we use multiple sequence alignments in our experiment?

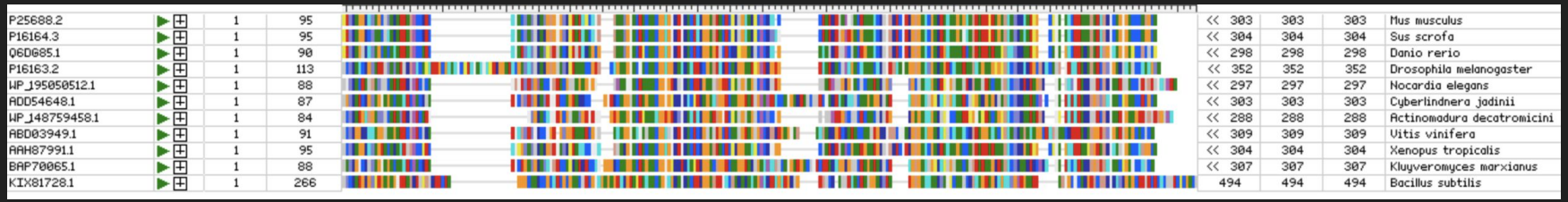
1. We compared urate oxidase sequences from various species to choose our original mutation during our grant proposal draft phase of the experiment.

2. We also performed a multiple sequence analysis on the sequences of various SARS-CoV-2 main protease isoforms for one of the additional questions of our first lab report

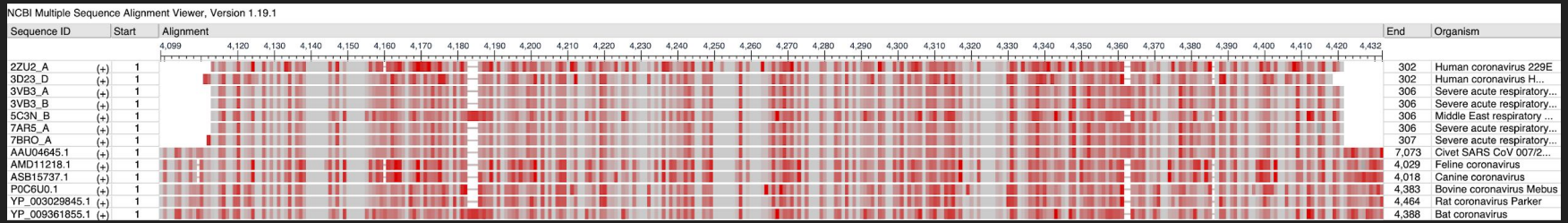
We could also have used a sequence alignment for our plasmid, which we sent in for sequencing. This would have allowed us to check for our mutation and any other errors that may have occurred.

I obtained the following MSAs

1. MSA of urate oxidase isoforms:



2. MSA of SARS-CoV-2 isoforms:



How I obtained these alignments

For the first MSA, I only used the MSA viewer in the Genome Workbench.

For the second, I used ESPript 3.0 to clean up the exported MSA file from genome workbench.

ESPript 3.0 VS 

Questions & Answers

Do you have any questions about how the MSA algorithm works?

Any specific questions regarding how to use the programs?

Anything else about my experiment?

References

1. Rosenberg, M. S. In *Sequence Alignment: Methods, Models, Concepts, and Strategies*; University of California Press, 2009; pp 3–21.
2. Higgins, D. G.; Sharp, P. M. CLUSTAL: a Package for Performing Multiple Sequence Alignment on a Microcomputer. *Gene* **1988**, *73* (1), 237–244.
3. Lipman, D. J.; Altschul, S. F.; Kececioglu, J. D. A Tool for Multiple Sequence Alignment. *Proceedings of the National Academy of Sciences* **1989**, *86* (12), 4412–4415.
4. Yainoy, S.; Phuadraksa, T.; Wichit, S.; Sompoppokakul, M.; Songtawee, N.; Prachayasittikul, V.; Isarankura-Na-Ayudhya, C. Production and Characterization of Recombinant Wild Type Uricase from Indonesian Coelacanth (L. Menadoensis) and Improvement of Its Thermostability by In Silico Rational Design and Disulphide Bridges Engineering. *International Journal of Molecular Sciences* **2019**, *20* (6), 1269.

Shane Kalani Abbley

Chem 125L - Kalju Kahn