README FILE – Nucleotide Sequence Comparison Tool

This is a program written in Python to perform a comparative genomic analyses of two sets of genomic sequence data to be specified by the user. It is designed to output to Excel file, the sequence positions where nucleotide residues are equal, where they are different (and what the difference is) and where there are gaps in either sequence with respect to the other. It was originally intended to compare sequence data from clones of the aphid *M. persicae*, but has been modified for genomic sequence data belonging to the *Barley yellow dwarf virus* (an aphid borne crop virus).

It currently requires pre-aligned genomic files of equal length as inputs, to facilitate pairwise comparison between nucleotides at the correct positional indices in the genomic sequences. Similarly, input files should contain only an unbroken string of nucleotides and “-“ characters to represent gaps in either of the input sequences.

This program is applicable to genomic sequence data from other organisms, however will require formatting of the input files as mentioned above. The file names and file handles can also be changed by the user to reflect the chosen source genomic sequence data.

To run:

1. Download the Python scripts from this repository, together with the associated input data files. A Jupyter notebook file is also available to explain how the script works and its design process.
2. The relevant input file names should already be present in the script(s), so modify if you wish to use another dataset prepared according to the above criteria.
3. Two variants of the script are available. The first of these outputs all index positions and the sequence identity between nucleotide sequences as 0%, 100% or “Gap” (Complete nucleotide sequence identity script.py). The second returns only the differences between nucleotides (0%) and missing nucleotide (“Gap”) indices. (Nucleotide sequence identity – differences only script.py)
4. The script may need to be run with python3 command from terminal on certain computers.
5. The pandas data analysis library should be installed in addition to Python before running scripts.
6. Output files from either script are also provided in the repository as an example.