SEQUENCE ALIGNMENT INTRODUCTION USING CLUSTALX

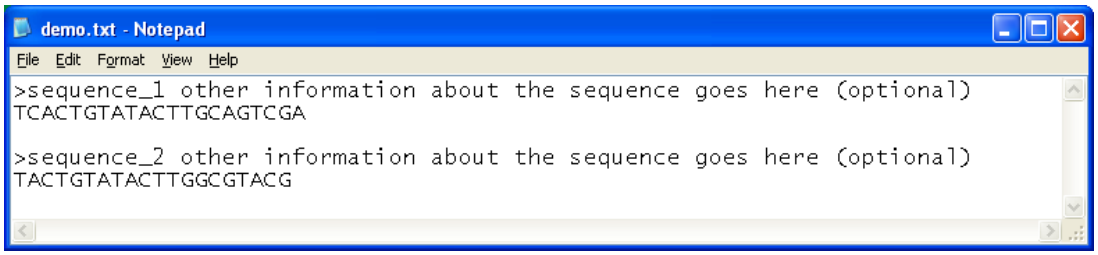
This document can be used to introduce the basic concept of DNA sequence alignment, which is necessary before DNA sequences can be meaningfully compared.

FORMAT OF DNA SEQUENCE INFORMATION

There are different formats for representing DNA sequences. Shown below is a partial sequence from the dog's *cytochrome oxidase subunit I* (*COI*) gene in FASTA format. FASTA format starts with a ">," followed by information about the file to the end of the first line, followed by the DNA sequence.

> >gi|377685879|gb|JN850779.1| Canis lupus familiaris isolate dog_3 cytochrome oxidase subunit I (COI) gene, partial cds; mitochondrial
>
> TACTTTATACTTACTATTTGGAGCATGAGCCGGTATAGTAGGCACTGCCTTGAGCCTCCTCATCCGAGCC
>
> GAACTAGGTCAGCCCGGTACTTTACTAGGTGACGATCAAATTTATAATGTCATYGTAACCGCCCATGCTT…

A file containing FASTA format sequence information may contain multiple sequences one after another. For example:



WHAT SEQUENCES DO WE CHOOSE TO COMPARE?

In modern taxonomic practice, scientists routinely analyze the DNA from specimens they collect to obtain a "DNA barcode," a short DNA sequence unique to a particular species, which is used to identify the species it belongs to. For animals and many other eukaryotes, different genes have been used for this purpose. One example is the mitochondrial cytochrome oxidase subunit I (*COI*) gene, which encodes part of an enzyme that is important for cellular respiration, and the mitochondrial NADH dehydrogenase subunit 2 (*ND2*) gene is another. Sequences like these are available from a wide range of species, making it possible to use these gene sequences to explore phylogenetic relationships.

COI or *ND2* are good choices for DNA barcoding, because, in general, there is little variation in the sequences of organisms within the same species, while there is significant variation in the sequences of organisms from different species. Therefore, they provide a unique sequence signature for a particular species, and are suitable for comparing phylogenetic relationships between species.

Because these sequences are so similar within the same species, these genes are not a good choice for studying variations within the same species, or even among species that have very recently speciated.

COI sequences also have a low mutation rate among many species of plants and cannot be used for DNA barcoding or phylogenetic comparisons of those species.

The sequence includes the NADH dehydrogenase subunit 2 gene along with adjacent sequences that include some transfer RNA genes. ND2 gene is one of several genes that are often used for genetic fingerprinting in animals. It is suitable for this purpose because it is conserved enough so that the gene is shared among a diverse group of animals, yet different enough between different animals to examine evolutionary relationship by comparing DNA sequences.

WHICH PROGRAM TO USE

To teach the basics of sequence alignment, we recommend ClustalX if you can install software on your computer. To generate phylogeny, using www.phylogeny.fr is simpler.

- ClustalX has a graphic interface that is intuitive, and it is an excellent tool for illustrating the concept and the process of sequence alignment. ClustalX is a freely available installed program, with its advantage (no reliance on internet) and disadvantage (requires program installation) in the classroom setting. Its algorithm is also a little dated, and there are other programs that do a better job of generating phylogenies; however it is sufficient as a demonstration of how to generate phylogenetic trees from DNA sequence alignments. The phylogeny generated requires another freely available program, NJplot, to print or view.
- www.Phylogeny.fr is a web-based tool for generating phylogenies. Using the default settings, phylogeny.fr is simple to use, andit uses a different alignment generator called MUSCLE. The website generates a phylogeny that can be saved as different graphic files. However, the display of alignment is not as intuitive as in ClustalX.

ALIGNMENT TUTORIAL AND TREE GENERATION VIA CLUSTALX

Software and Files

Install ClustalX, which is available at http://www.clustal.org/clustal2/. (For Windows, download clustalx-2.1-win.msi; for Mac OS, download clustalx-2.1-macosx.dmg.) Next, install NJplot, which is available at http://pbil.univ-lyon1.fr/software/njplot.html.

Understanding sequence alignment

Let's use ClustalX to compare DNA sequences. For this exercise, use the test sequence file test.txt, which contains the three short DNA sequences (test1, test2, and test3) shown below:

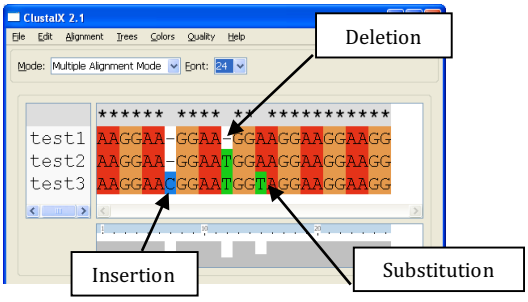
>test1
AAGGAAGGAAGGAAGGAAGGAAGG
>test2
AAGGAAGGAATGGAAGGAAGGAAGG
>test3
AAGGAACGGAATGGTAGGAAGGAAGG

Load these sequences into ClustalX by choosing from the menu, File -> Load Sequences, and then selecting test.txt. ClustalX displays these sequences as shown in the illustration on the right (PC version shown).

Before you can compare sequences, you have to "align" them, which means lining up the sequences and sliding them past one another until the best matching pattern is found. Alignment allows you to examine differences between related sequences; such differences reflect evolutionary relationships.

From the menu, choose Alignment -> Do Complete Alignment. When prompted for output file names, use the default names given and click "OK." The screen changes to looks like the illustration on the right.

Notice that it's a lot easier to see differences among DNA sequences after alignment. You can figure out what kinds of mutations have occurred in each sequence by how it compares to the others (as shown in the labeled illustration). The number of differences among sequences determines how closely or distantly related the corresponding organisms are.



Based on this information alone, which two sequences do you think are more closely related? To see if your answer was accurate, we can use ClustalX to generate a phylogenetic tree.

From the menu, choose Trees -> Draw Tree. This creates a phylogenetic tree file called test1.ph, which can be opened using NJplot.exe. **Launch NJplot, then from its menu, choose File -> Open, and select test1.ph.**

The result shows that test1 and test2 are on the same branch of the tree, indicating that they are more closely related to each other than to test3.