

# Bioinformatics: Searching Online Databases for DNA Sequences

---

## Learning Objectives

After completion of this module, the student will be able to

- search for sequence data using online public databases
- using online alignment tools to align multiple sequences
- appreciate the large amount of DNA and protein sequence data that are publicly available
- interpret text and visual representations of relatedness of sequences

## Knowledge and Skills

- Dot plot
- Local sequence alignment and scoring
- Introduction to BLAST
- Tree building using BLAST resources

## Prerequisites

- Introduction to DNA sequences

## Introduction

In the last 10 to 15 years, the amount of sequence data that is available has grown, and continues to grow, exponentially. Most of the sequence data is publicly available. As sequences are discovered they are typically placed in online databases where anyone can access them. The open sharing of knowledge can lead to discoveries that would never have been found since you may have a different reason for using the data than the person who originally obtained the sequence.

A common task in bioinformatics is the alignment of two or more DNA or protein sequences. There are several reasons why you may want to do this. You may have sequenced a gene of unknown function and might want to search a database for similar sequences to help you in identifying its function. You may want to see how similar two sequences are and estimate how long ago they diverged. Alignment of genes is a common first step in other processes such as building a phylogeny.

## Pairwise Sequence Alignment

We will begin by aligning two sequences at a time. This is called *pairwise sequence alignment*, as opposed to *multiple sequence alignment*.

If large segments of two sequences are identical, they may be easy to align. Look at the following example:

```
G A A T G C A A
T G A A T G C A
```

After a quick glance you may see that the sequences align well in the following way:

```
  G A A T G C A A
T  G A A T G C A
```

However, how do you know for sure that this is the best alignment? How can you define “best alignment” so people working independently can agree on the same result?

## Dot Plots

A straightforward but powerful way to analysis an alignment is to use a dot plot. A dot plot is a matrix that is used to visually detect alignments. To illustrate how dot plots work, we use the same sequences as above. We begin making our dot plot by putting the first sequence on top and the second one on the left side. Orientation does not matter. For example, we could have put the second sequence on the bottom and the first one on the right side. But let’s stick with putting one sequence on top and the other on the left side:

	G	A	A	T	G	C	A	A
T								
G								
A								
A								
T								
G								
C								
A								

Next, find the cells that correspond to the same type of nucleotide in both sequences. For instance, the first cell in the second row corresponds to a G in the first position in the top sequence and a G in the second position in the sequence on the left side. We mark the cell with an “X”:

	G	A	A	T	G	C	A	A
T								
G	X							
A								
A								
T								
G								
C								
A								

Using an X works well, but you can mark the cells however you want. You can color the whole cell or insert a letter. Any symbol is fine. Traditionally, the whole cells were shaded black. When using sequences that are thousands of base pairs long, the matrix then looks like a plot of dots, hence the name “dot plot.” Filling in all the matches, we find

	G	A	A	T	G	C	A	A
T				X				
G	X				X			
A		X	X				X	X
A		X	X				X	X
T				X				
G	X				X			
C						X		
A		X	X				X	X

At this point, visualizing the best alignment may still be difficult. To make patterns easier to see, we search for any diagonal, consecutive series of marked cells. These marked cells form *tuples*. Tuples are also called *words*. We will require that the number of cells in a series that forms a tuple must be equal to or greater than a certain *word size* in order to be highlighted. You will specify which word size to use. Let’s use a word size of 4. We draw a line through the words that are at least 4 cells in length. Next we unmark all the tuples that do not meet the required word length. We obtain

	G	A	A	T	G	C	A	A
T								
G	X							
A		X						
A			X					
T				X				
G					X			
C						X		
A							X	

In this example, there is only one word, or tuple, that meets our requirement. The tuple in the dot plot corresponds to the following alignment:

```

      G A A T G C A A
T    G A A T G C A
    
```

Now let's compare the first sequence to a third sequence:

```

      G A A T G C A A
T    G C A T G C A
    
```

How do we decide what the best alignment is? Like earlier, we can use a dot plot, and let's use again a word size of 4. We first mark matching cells:

	G	A	A	T	G	C	A	A
T				X				
G	X				X			
C						X		
A		X	X				X	X
T				X				
G	X				X			
C						X		
A		X	X				X	X

We then identify words of length 4 or more and unmark all words that do not meet the requirement:

	G	A	A	T	G	C	A	A
T				X				
G					X			
C						X		
A		X	X				X	X
T				X				
G	X				X			
C						X		
A		X	X				X	X

We now have two tuples corresponding to two alignments:

```

      G A A T G C A A
      T G C A T G C A

      G A A T G C A A
T    G C A T G C A
    
```

To decide which alignment is better, we need a way to score alignments.

### Scoring Algorithms

One basic method is to slide one sequence against the other, give each alignment a score, and then pick the one with the best score. This way of aligning only looks at matches and mismatches. It ignores insertions and deletions. To score an alignment, a *substitution matrix* is used. The substitution matrix tells you what score to give each pairs of bases in an alignment. The simplest substitution matrix gives each identical pair of nucleotides a score of 1 and each pair of different nucleotides a 0. The substitution matrix is the identity matrix. Adding up the score for each aligned pair gives the alignment score. Let's do this for the same sequences we just used to create a dot plot.

```
  G A A T G C A A
  T G C A T G C A
```

Using this algorithm, we can score the two alignments we found using the dot plot:

```
      1 1 1 1 0
G  A  A  T  G  C  A  A
      T  G  C  A  T  G  C  A

      1 0 1 1 1 1 1
G  A  A  T  G  C  A  A
T  G  C  A  T  G  C  A
```

Adding up the individual scores, we find that the first alignment has an alignment score of 4 and the second one has an alignment score of 6. Since the second alignment has the higher score, 6 vs. 4, we choose the second alignment as the better one. The mismatch in the second alignment (highlighted in red) was likely a point mutation at some point along one of the two lineages. Since we do not know the ancestral sequence, we do not know if an A mutated into a C or a C mutated into an A, or both sites experienced mutations.

### BLAST

Doing alignments by hand only works for very short sequences. There are online resources available to find alignments. BLAST is probably the most widely used online program for making pairwise alignments. BLAST is an acronym for *Basic Local Alignment Search Tool*, and its primary purpose is to find genes in a database that have similar sequences.

The BLAST algorithm is an example of a heuristic method. It is quite a bit more complicated than the simple scoring algorithm that we introduced above. In very general terms, BLAST takes one of the sequences and splits it up into words, or tuples. You can decide how big to make the words. The default for nucleotides is a size of 11. For amino acids the default word size is 3. For example, the amino acid sequence MPQEG can be divided into 3 tuples each with size 3: MPQ, PQE, and QEG. Each word is then slid against the other sequence. Each alignment is given a score. If the score is above a specified threshold, the alignment is retained. Each local alignment is then extended and measured using something called a high-scoring segment pair (HSP) score. If there are two or more HSP regions, they will be combined, if possible. If it is not possible, you may end up with multiple HSPs. For more details, see the following websites:

[http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/BLAST\\_algorithm.html](http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/BLAST_algorithm.html)

<http://en.wikipedia.org/wiki/BLAST>

[http://blast.ncbi.nlm.nih.gov/blast/blast\\_help.shtml](http://blast.ncbi.nlm.nih.gov/blast/blast_help.shtml)

### **Use BLAST to Search Databases**

We will use BLAST to find genes in a database that have similar sequences. Why would you want to search a database for similar genes? Perhaps you found a gene that causes a disease in humans and you want to find a gene with the same function in a model organism so you can do genetic studies. Or perhaps you sequenced a gene of unknown function and want to find a similar gene of known function in some other organism to infer the function of the unknown gene. Or perhaps you may be interested in understanding the evolution of a gene family and want to create a phylogeny of the gene family using genes from multiple species. Using BLAST is an easy way to search a large database for the genes you need.

We will use BLAST to search the Microbes database to find closely related organisms for an unknown ancient microbial DNA sequence. The DNA sequence that forms the basis of the search is called the *query sequence*.

Now go to <http://blast.ncbi.nlm.nih.gov/Blast.cgi> We will search the Microbes BLAST database for genes with sequences similar to our query sequence. Select the Microbes database.

**BLAST** Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

NCBI/BLAST Home

BLAST finds regions of similarity between biological sequences. [more...](#)

**New** Designing or Testing PCR Primers? Try your search in **Primer-BLAST**.

### BLAST Assembled Genomes

Choose a species genome to search, or [list all genomic BLAST databases](#).

- [Human](#)
- [Mouse](#)
- [Rat](#)
- [Arabidopsis thaliana](#)
- [Oryza sativa](#)
- [Bos taurus](#)
- [Danio rerio](#)
- [Drosophila melanogaster](#)
- [Gallus gallus](#)
- [Pan troglodytes](#)
- [Microbes](#)
- [Apis mellifera](#)

### Basic BLAST

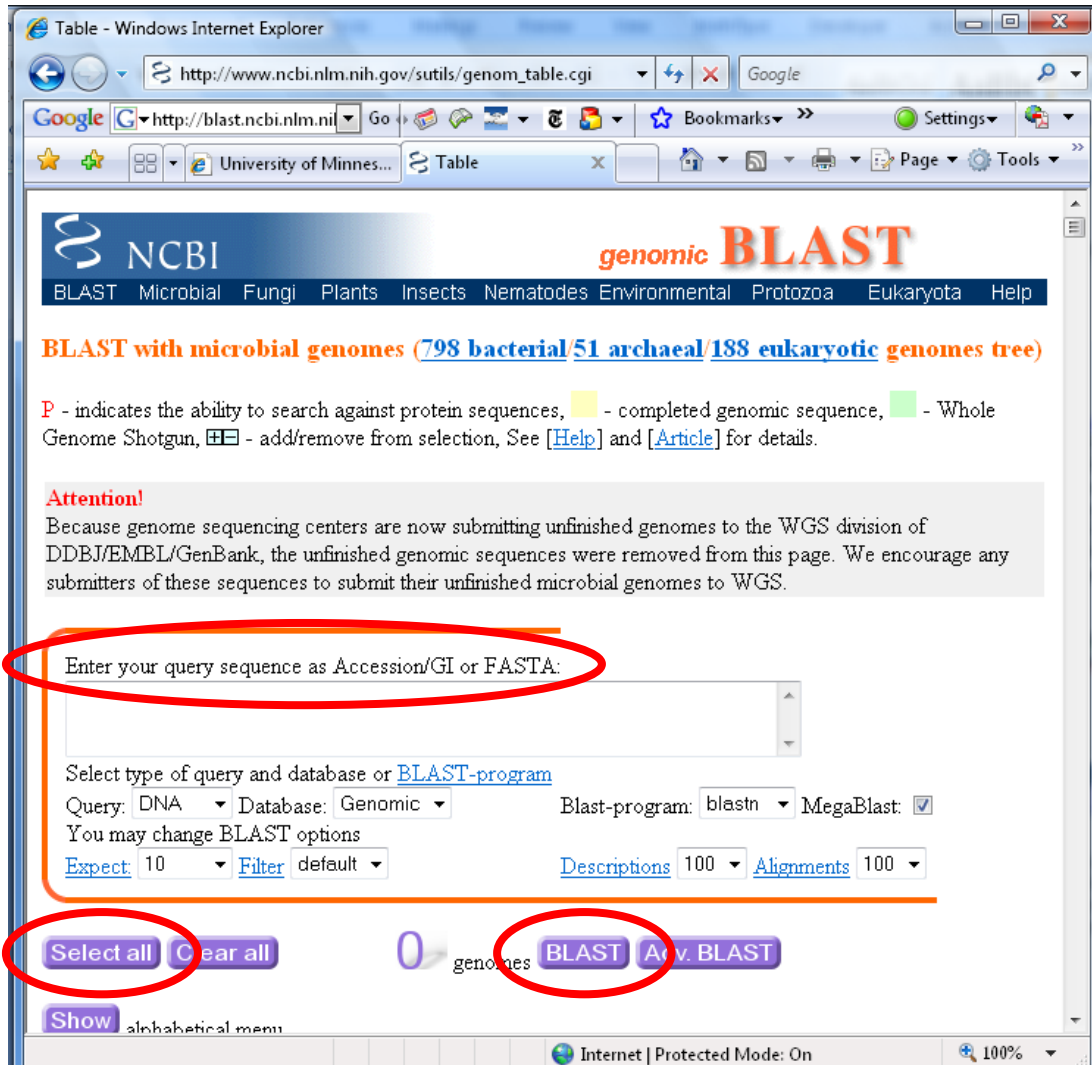
Choose a BLAST program to run.

<a href="#">nucleotide blast</a>	Search a <b>nucleotide</b> database using a <b>nucleotide</b> query <i>Algorithms: blastn, megablast, discontinuous megablast</i>
<a href="#">protein blast</a>	Search <b>protein</b> database using a <b>protein</b> query <i>Algorithms: blastp, psi-blast, phi-blast</i>
<a href="#">blastx</a>	Search <b>protein</b> database using a <b>translated nucleotide</b> query

The query sequence for our case study is the ancient DNA sequence that was extracted from the dental pulp of teeth found in the grave that dates back to the Peloponnesian war (Papagrigrakis et al. 2006).

```
GTTCACTTCTGCCATGAGGAGCGAACAAAACCACCACGACCACGCGCCTGTTTGAAGCTTTTGGCTTTATCGGCA  
TCTTCAATGATGGAGGCCCATGCCTTACCGGATCGCGGTGCAGTTTCTTGGCCTCGCGCCACAGTTTGATCAGGC  
GTTTGCGCATCAGCGGGTATTTTCAGGCGGTTGGCGCTGTACAGATACCAGGAGTAACTGGCGCCACGTGGGCAAC  
CGAGCCGTTTCATGGTTGGGCAAGTCCGGGCGGGTACGCGGATAATCGGTTTGCTGGGTTTCCCAGGTCACCAGGC  
CATTTTTACATAAATTTCCAGCTACAGGAGCCGGTGCAGTTCACCCCATGGGTTGA
```

Paste the nucleotide sequence of the query sequence into the box under "Enter your query sequence..."  
Click on "Select all" and then "BLAST."



This initial search will give you an idea of which microbial DNA sequences are closely related to the ancient DNA. Take it from there!

## References

Papagrigoakis, M.J., C. Yapijakis, P. N. Synodinos, and E. Baziotopoulou-Valvani. 2006. DNA examination of ancient pulp incriminates typhoid fever as a probable cause of the Plague of Athens. *International Journal of Infectious Diseases* 10: 206-214.