Bioinformatics:

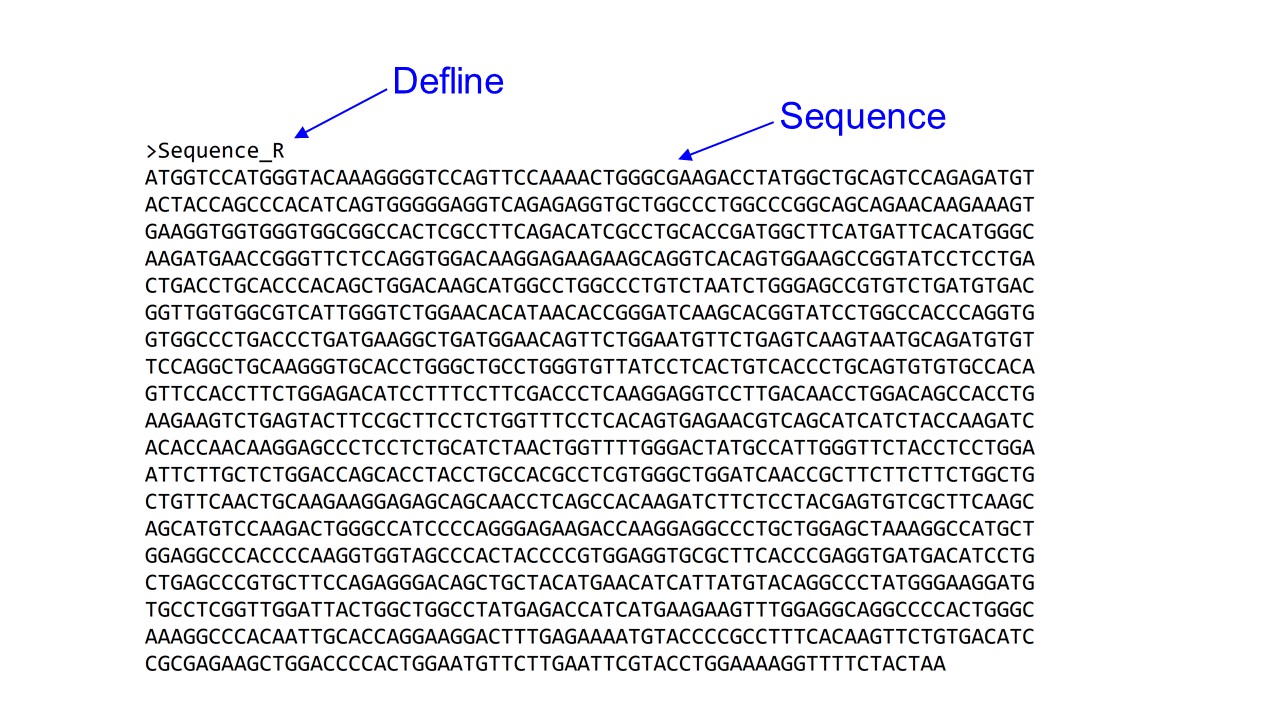
The Power of Computers in Biology

Worksheet

**Task A: Identify the Mysterious “Nucleotide Sequence R”**

|  |
| --- |
| We provide a DNA sequence without any information on its function – **Sequence R**. |
|  |
| 1. Open a Web browser. |
|  |
| 1. Go to our Website [www.4273pi.org](http://www.4273pi.org) and click the “STUDENTS” tab. You will find a link to **Sequence R**.  Copy the whole of it (including the defline “>Sequence\_R”). |
|  |
| **Keep this browser window open**. |

The sequence is laid out in *Fasta format* (Figure 1).



**Figure 1.** Components of a sequence record in Fasta format.

To discover the role of this mysterious sequence, we will search for proteins in the database that show high similarity to a translation of **Sequence R**.

To do this we will run BLAST at the National Center for Biotechnology Information (NCBI). The NCBI maintains a huge database of known DNA and protein sequences.

|  |
| --- |
| 1. Open a new tab in your Web browser and search for “NCBI”. Click the link to “National Center for Biotechnology Information”. |
| 1. Under “Popular Resources” on the right, click “BLAST”, then “blastx”. |
| 1. In the Web browser, in the box labelled “Enter Query sequence”, paste **Sequence R**. |
| 1. Click “BLAST”. |

BLASTX uses the genetic code to translate **Sequence R**, then compares it with every protein in the sequence database.

The BLASTX search may take a few minutes, during which time “Status” on the Web page is “Searching”. Results will be shown in a long Web page, headed “BLAST >> blastx >> results for ...” Scroll down to see a table of results. The best-matching sequence from the database is listed first.

|  |
| --- |
| QUESTION 1.  **Sequence R** has an excellent match to a known protein, indicated by its low E-value and high percentage identity. We assume **Sequence R** codes for this protein.  What is the name of this protein? |
|  |
| ANSWER: |

|  |
| --- |
| QUESTION 2. In which organism is this protein found? |
| **Hint**: *Click the link under “Accession” to go to the database record for this protein. Look for the line beginning “SOURCE”.* |
|  |
| ANSWER: |

|  |
| --- |
| **QUESTION 3**. What is the biological role of the protein? |
|  |
| ***Hint****: Do a Web search for the name of the protein (from your answer to Question 1).* |
|  |
| ANSWER: |

**Task B: Search for a match to Sequence R in the human genome**

We will now perform another blast search to see if Sequence R has a match in the human genome. We will now use BLASTN, which compares a DNA query with a DNA sequence database.

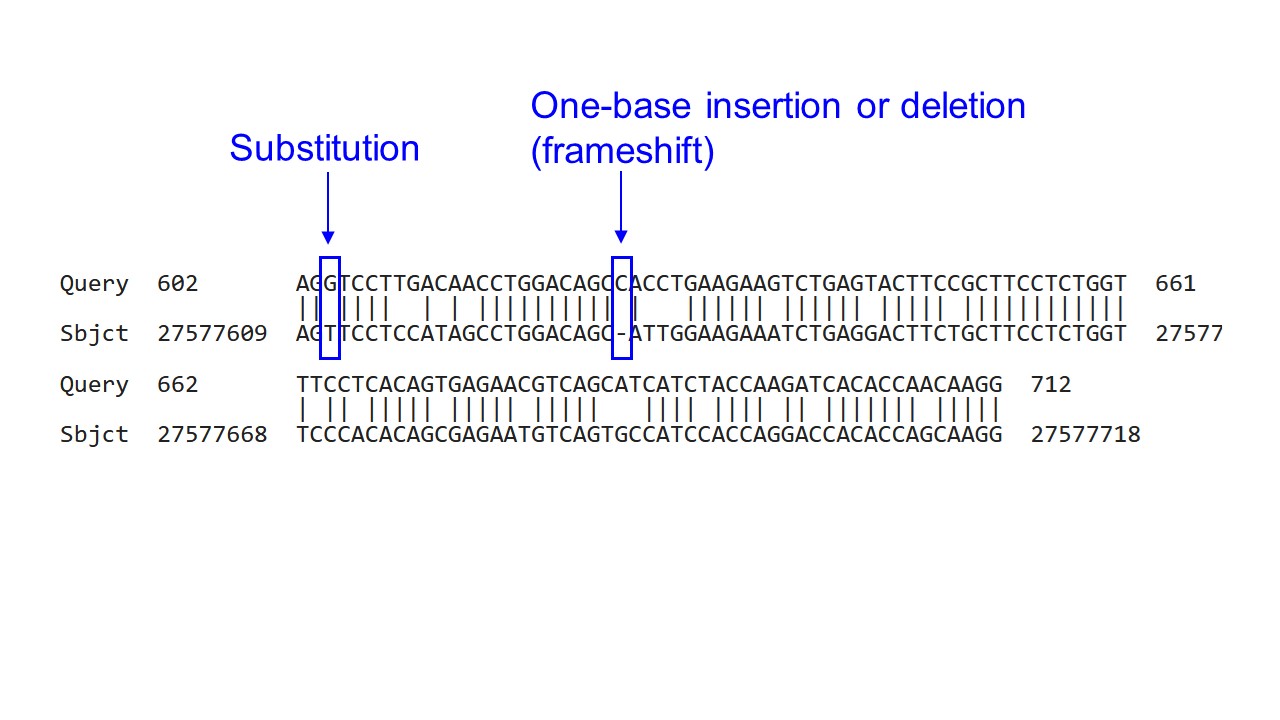
|  |
| --- |
| 1. Copy **Sequence R** again from our Web page. |
| 1. Go back to the main NCBI BLAST Web page. (If you have closed the tab, you can just search for “NCBI BLAST”.) Under “BLAST Genomes”, click “Human”. |
|  |
| 1. Click “blastn” at the top of the page, and paste in **Sequence R**. |
| 1. Under “Database”, choose “Genome (GRCh38.p14 reference assembly)” |
| 1. Under “Optimize for”, select “Somewhat similar sequences (blastn)”. |
|  |
| 1. Run a BLASTN search with **Sequence R** as query. |

|  |
| --- |
| QUESTION 4.  On which human chromosome has the best match been found? |
|  |
| **ANSWER:** |
|  |

|  |
| --- |
| Click on the first result in the ‘Description’ column of the table. This will display fragments of **Sequence R** (the query sequence) aligned against the fragments of human genomic DNA found by the BLASTN search (the subject sequences).  Look at the BLASTN results carefully. |

In your results, you will see a good match between **Sequence R** and the human genomic DNA. However, the mouse and human sequences differ due to mutations.

Figure 2 (on the next page) shows evidence of substitution and frameshift mutations in BLASTN results.



**Figure 2.** Finding mutations in BLASTN results.

|  |
| --- |
| QUESTION 5.  In the alignment between **Sequence R** and the human genomic DNA, can you see evidence of a substitution mutation? If so: sketch the region that includes the substitution. |
|  |
| ANSWER: |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |

|  |
| --- |
| QUESTION 6.  Can you see evidence of an insertion or deletion mutation? If so: sketch the region that includes the insertion or deletion. |
|  |
| ANSWER: |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |

A **frameshift mutation** is an insertion or deletion which disrupts the reading frame of a protein-coding sequence. From this point on, any protein sequence would be scrambled.

A frameshift mutation is strong evidence that the DNA no longer codes for a functional protein. **These “former genes” are called *pseudogenes*.**

|  |
| --- |
| QUESTION 7.  Do you think the human genome includes a functional version of **Sequence R** or is there a pseudogene instead? Explain your answer. |
|  |
| ANSWER: |
|  |
|  |
|  |
|  |
|  |

|  |
| --- |
| QUESTION 8. |
|  |
| **Sequence R** comes from the house mouse and codes for L-gulonolactone oxidase, an enzyme that synthesizes vitamin C. |
| Vitamin C is vital for both humans and mice. |
|  |
| Does your answer to Question 8 tell us anything about how the diet of humans might differ from the diet of mice? |
|  |
|  |
| ANSWER: |
|  |
|  |
|  |
|  |

Task C: Using BLAST at the command line

We typically interact with a computer using a graphical user interface (GUI). This involves using a mouse or a touch-screen, perhaps with some typing. GUIs can be very easy to use. However, GUIs can be inflexible.

An alternative is the **command line**. This involves commands typed into a terminal.

The command line may be more difficult to use at first. However, it is more flexible. Also, we can easily keep a copy of the commands we ran, making it easier to document our actions.

In bioinformatics, both GUI and command line are used every day. The most convenient command line for scientific computing is provided by a Linux operating system.

# Introduction to the command line: creating, saving, locating, reading and removing files

|  |
| --- |
| Follow our on-screen instructions to connect to the Linux server.  To access the command line, click the black rectangular icon near the top left. |

We will now first use the nano text editor to create a file.

|  |
| --- |
| To start editing a new file called test.txt, type the following text, then ENTER |
|  |
| *nano test.txt* |
|  |
| You now see something like a basic word processor. You can type text, move around with the arrow keys, and delete text. |
|  |
| Type some text, for example “hello world”, your name or favourite food. |
|  |
| To exit nano and save your file: |
|  |
| *Press CTRL-X* |
|  |
| nano will then ask: |
| *Save modified buffer (ANSWERING "No" WILL DESTROY CHANGES) ?* |
|  |
| Press y to save your file. |
|  |
| nano will then ask for the file name, suggesting the name you gave earlier (test.txt). |
|  |
| *Press ENTER.* |
|  |
| nano should now close, which brings you back to the terminal window. |

To find the file you created, we now will list all files in the current directory.

|  |
| --- |
| Type the following text, then ENTER |
|  |
| *ls* |
|  |
| **Hint***: this command is short for “list”, it starts with a lower case L, not the numeral “one”.* |

You should see the name of the file you just created – test.txt – and others. You can display the contents of a text file using the cat command.

|  |
| --- |
| Type the following text, then ENTER |
|  |
| *cat test.txt* |

When you are happy that your file contains the expected text, you can delete it. The command rm deletes a file permanently.

|  |
| --- |
| Type the following text, then ENTER |
|  |
| *rm test.txt* |
|  |
| To check that the file has been removed, type the following text, then ENTER |
|  |
| *ls* |
|  |
| If the file was removed, you should no longer see the file name test.txt displayed. |

# Sequence data in text files

The file sequence\_r.fa contains the same DNA sequence you used in Tasks A and B.

|  |
| --- |
| To display the contents of sequence\_r.fa , type the following text, then ENTER |
|  |
| *cat sequence\_r.fa* |
|  |
| **Tip:** It is time-consuming and error-prone to type file names. If you start typing and then press the TAB key, the command line will complete the filename itself, where possible.  So, you could save time by typing: |
|  |
| cat s followed immediately by pressing the TAB key |
|  |
| After pressing the TAB key, you should see the full file name appear: sequence\_r.fa |
|  |
|  |

The file Homo\_sapiens.GRCh38.dna.chromosome.8.fa contains the DNA sequence of human Chromosome 8, in Fasta format.

|  |
| --- |
| Type the following text, then ENTER |
|  |
| *cat Homo\_sapiens.GRCh38.dna.chromosome.8.fa* |
|  |
| **Hint:** *to save time and avoid typing errors, try typing H and then pressing the TAB key.* |
|  |
| This DNA sequence is very long. If it takes too long to display the whole sequence, you can top the process by pressing CTRL-C. |
|  |

# Running BLAST at the command line

Shape

Description automatically generated with medium confidence

|  |
| --- |
| Now, we will use **Sequence R** as a query in a BLAST search, with Chromosome 8 as database. |
|  |
| Type the following text (all in one line), then press ENTER |
|  |
|  |
| **Hint**: There is a single space between each word. Press ENTER at the end of the command (after results.txt). |

The meaning of this command is as follows:

|  |  |
| --- | --- |
| blastn | run BLASTN, which searches a DNA sequence file or database with a DNA sequence query. |
|  |  |
| -task blastn | equivalent to the “Optimize for somewhat similar sequences (blastn)” option when you used BLAST from the NCBI Web site. |
|  |  |
| -query sequence\_r.fa | use sequence\_r.fa as the query for the search. |
|  |  |
| -subject Homo\_sapiens.GRCh38.dna.  chromosome.8.fa | search for the query in the Homo\_sapiens.GRCh38.dna.chromosome.8.fa file. |
|  |  |
| > results.txt | send the output to a text file named results.txt. |

|  |
| --- |
| Now look at the results in nano. Type the following text, then ENTER |
|  |
| nano results.txt |
|  |
| **Tip**: *Increase the size of the terminal window, by clicking on an edge of the window with the mouse and dragging.* |
|  |
| In nano, you can scroll up and down one line at a time with the arrow keys, or one page at a time with CTRL-Y and CTRL-V |

|  |
| --- |
| QUESTION 9.  On the basis of the command-line BLAST search, do you think these results represent a functional gene in the human genome, or is there a pseudogene instead?  Is this the same conclusion you drew from the search you ran via the NCBI Web site? |
|  |
| ANSWER: |
|  |
|  |
|  |

# Documenting your work

A log of recent commands can be displayed using the history command. This can be very useful to track down mistakes, or to share your work with others.

|  |
| --- |
| Exit nano. In the terminal window, type the following text, then ENTER |
|  |
| *history* |

|  |
| --- |
| Worksheet Version 2.0  By Daniel Barker, Heleen Plaisier, Stevie A Bain and Richard Fitzpatrick |
| 4273pi Bioinformatics Education Project, <https://4273pi.org> |
|  |
| Copyright and related rights waived via CC0 1.0 Public Domain Dedication (<https://creativecommons.org/publicdomain/zero/1.0>). |