Introduction to Bioinformatics

Some expertise in bioinformatics is required for all students in life sciences. In this document we introduce Bioinformatics and provide you with useful tools.

The content of this document is linked to the practical exercise of this week, which deals with the regulation of the Lac operon in *Escherichia coli.* Read the text and carry out the assignments it contains. At the end a test allows you to verify your answers.

This Bioinformatics assignment must be completed by each student because its contents are part of the requirements for the final exam.

# What is bioinformatics?

Bioinformatics is the scientific discipline that deals with the organisation and interpretation of information present in DNA and protein sequences.

Bioinformatics is used in life sciences to tackle biological problems or develop new medical approaches, such as evolution of species or identification of genetic lesions in cancer cells.

DNA sequences and amino acid sequences from various organisms are stored in online databases. These databases currently contain hundreds of billions of nucleotides and this number continuously increases.

In this assignment you will use Bioinformatics tools:

1. To locate sequences in DNA that play a role in gene expression.

2. To identify a gene based on its sequence.

# Finding genes in a DNA sequence

Page 5 shows 6400 nucleotides of the *Escherichia coli* genome. The entire genome of *E. coli* is about 750 times larger, and a eukaryotic genome is generally more than a million times longer.

The DNA sequence on page 5 starts at the 5’- side and ends at the 3’- end. **Only one of the DNA strands is shown. The sequence of the complementary strand is in reverse. Both strands can encode genes, which are always transcribed from the 5’ to 3’ direction.**

To determine which genes are present in a DNA sequence, you should first understand how nucleotides code for proteins. Amino acids, the building blocks of proteins, are programmed in codons. A codon is a group of 3 consecutive nucleotides, which are also called triplets. Codons can be translated into amino acids, except for the stop codons (see below figure (Fig. 7-27, Essential Cell Biology, fifth edition)).

A picture containing table

Description automatically generated

## Six reading frames exist in DNA

Reading frame 1 starts at the first nucleotide in the DNA sequence on page 5: GCG-CAA-AAC-…, which encodes for the amino acid sequence Alanine-Glutamine-Asparagine-…. (see table 2.1).

You could also start reading at the second nucleotide. The triplet-sequence is then CGC-AAA-ACC-…, encoding for Arginine-Lysine-Threonine-… (this we call reading frame 2). Beginning at the third nucleotide the DNA triplet-sequence is GCA-AAA-CCT-…, which encodes Alanine-Lysine-Proline-… (i.e., reading frame 3).

Note that the reverse complementary strand of the sequence on page 5 also contains three reading frames. Therefore, six reading frames exist in DNA.

Table 2.1: Three reading frames can be distinguished in a DNA sequence

|  |  |  |
| --- | --- | --- |
| Frame | Sequence | |
| 1 | DNA:  Translation: | **G C G|C A A|A A C|C T.**  **Ala – Gln – Asn** |
| 2 | DNA:  Translation: | **G|C G C|A A A|A C C|T...**  **Arg – Lys – Thr** |
| 3 | DNA:  Translation: | **G C|G C A|A A A|C C T...**  **Ala – Lys - Pro** |

### 6400 nucleotides of the *E. coli* genome

Use the sequence below to carry out the assignments.

GCGCAAAACCTTTCGCGGTATGGCATGATAGCGCCCGGAAGAGAGTCAATTCAGGGTGGTGAATATGAAACCAGTAACGTTATACGATGTCGCAGAGTATGCCGGTGTCTCTTATCAGACCGTTTCCCGCGTGGTGAACCAGGCCAGCCACGTTTCTGCGAAAACGCGGGAAAAAGTGGAAGCGGCGATGGCGGAGCTGAATTACATTCCCAACCGCGTGGCACAACAACTGGCGGGCAAACAGTCGTTGCTGATTGGCGTTGCCACCTCCAGTCTGGCCCTGCACGCGCCGTCGCAAATTGTCGCGGCGATTAAATCTCGCGCCGATCAACTGGGTGCCAGCGTGGTGGTGTCGATGGTAGAACGAAGCGGCGTCGAAGCCTGTAAAGCGGCGGTGCACAATCTTCTCGCGCAACGCGTCAGTGGGCTGATCATTAACTATCCGCTGGATGACCAGGATGCCATTGCTGTGGAAGCTGCCTGCACTAATGTTCCGGCGTTATTTCTTGATGTCTCTGACCAGACACCCATCAACAGTATTATTTTCTCCCATGAAGACGGTACGCGACTGGGCGTGGAGCATCTGGTCGCATTGGGTCACCAGCAAATCGCGCTGTTAGCGGGCCCATTAAGTTCTGTCTCGGCGCGTCTGCGTCTGGCTGGCTGGCATAAATATCTCACTCGCAATCAAATTCAGCCGATAGCGGAACGGGAAGGCGACTGGAGTGCCATGTCCGGTTTTCAACAAACCATGCAAATGCTGAATGAGGGCATCGTTCCCACTGCGATGCTGGTTGCCAACGATCAGATGGCGCTGGGCGCAATGCGCGCCATTACCGAGTCCGGGCTGCGCGTTGGTGCGGATATCTCGGTAGTGGGATACGACGATACCGAAGACAGCTCATGTTATATCCCGCCGTTAACCACCATCAAACAGGATTTTCGCCTGCTGGGGCAAACCAGCGTGGACCGCTTGCTGCAACTCTCTCAGGGCCAGGCGGTGAAGGGCAATCAGCTGTTGCCCGTCTCACTGGTGAAAAGAAAAACCACCCTGGCGCCCAATACGCAAACCGCCTCTCCCCGCGCGTTGGCCGATTCATTAATGCAGCTGGCACGACAGGTTTCCCGACTGGAAAGCGGGCAGTGAGCGCAACGCAATTAATGTGAGTTAGCTCACTCATTAGGCACCCCAGGCTTTACACTTTATGCTTCCGGCTCGTATGTTGTGTGGAATTGTGAGCGGATAACAATTTCACACAGGAAACAGCTATGACCATGATTACGGATTCACTGGCCGTCGTTTTACAACGTCGTGACTGGGAAAACCCTGGCGTTACCCAACTTAATCGCCTTGCAGCACATCCCCCTTTCGCCAGCTGGCGTAATAGCGAAGAGGCCCGCACCGATCGCCCTTCCCAACAGTTGCGCAGCCTGAATGGCGAATGGCGCTTTGCCTGGTTTCCGGCACCAGAAGCGGTGCCGGAAAGCTGGCTGGAGTGCGATCTTCCTGAGGCCGATACTGTCGTCGTCCCCTCAAACTGGCAGATGCACGGTTACGATGCGCCCATCTACACCAACGTGACCTATCCCATTACGGTCAATCCGCCGTTTGTTCCCACGGAGAATCCGACGGGTTGTTACTCGCTCACATTTAATGTTGATGAAAGCTGGCTACAGGAAGGCCAGACGCGAATTATTTTTGATGGCGTTAACTCGGCGTTTCATCTGTGGTGCAACGGGCGCTGGGTCGGTTACGGCCAGGACAGTCGTTTGCCGTCTGAATTTGACCTGAGCGCATTTTTACGCGCCGGAGAAAACCGCCTCGCGGTGATGGTGCTGCGCTGGAGTGACGGCAGTTATCTGGAAGATCAGGATATGTGGCGGATGAGCGGCATTTTCCGTGACGTCTCGTTGCTGCATAAACCGACTACACAAATCAGCGATTTCCATGTTGCCACTCGCTTTAATGATGATTTCAGCCGCGCTGTACTGGAGGCTGAAGTTCAGATGTGCGGCGAGTTGCGTGACTACCTACGGGTAACAGTTTCTTTATGGCAGGGTGAAACGCAGGTCGCCAGCGGCACCGCGCCTTTCGGCGGTGAAATTATCGATGAGCGTGGTGGTTATGCCGATCGCGTCACACTACGTCTGAACGTCGAAAACCCGAAACTGTGGAGCGCCGAAATCCCGAATCTCTATCGTGCGGTGGTTGAACTGCACACCGCCGACGGCACGCTGATTGAAGCAGAAGCCTGCGATGTCGGTTTCCGCGAGGTGCGGATTGAAAATGGTCTGCTGCTGCTGAACGGCAAGCCGTTGCTGATTCGAGGCGTTAACCGTCACGAGCATCATCCTCTGCATGGTCAGGTCATGGATGAGCAGACGATGGTGCAGGATATCCTGCTGATGAAGCAGAACAACTTTAACGCCGTGCGCTGTTCGCATTATCCGAACCATCCGCTGTGGTACACGCTGTGCGACCGCTACGGCCTGTATGTGGTGGATGAAGCCAATATTGAAACCCACGGCATGGTGCCAATGAATCGTCTGACCGATGATCCGCGCTGGCTACCGGCGATGAGCGAACGCGTAACGCGAATGGTGCAGCGCGATCGTAATCACCCGAGTGTGATCATCTGGTCGCTGGGGAATGAATCAGGCCACGGCGCTAATCACGACGCGCTGTATCGCTGGATCAAATCTGTCGATCCTTCCCGCCCGGTGCAGTATGAAGGCGGCGGAGCCGACACCACGGCCACCGATATTATTTGCCCGATGTACGCGCGCGTGGATGAAGACCAGCCCTTCCCGGCTGTGCCGAAATGGTCCATCAAAAAATGGCTTTCGCTACCTGGAGAGACGCGCCCGCTGATCCTTTGCGAATACGCCCACGCGATGGGTAACAGTCTTGGCGGTTTCGCTAAATACTGGCAGGCGTTTCGTCAGTATCCCCGTTTACAGGGCGGCTTCGTCTGGGACTGGGTGGATCAGTCGCTGATTAAATATGATGAAAACGGCAACCCGTGGTCGGCTTACGGCGGTGATTTTGGCGATACGCCGAACGATCGCCAGTTCTGTATGAACGGTCTGGTCTTTGCCGACCGCACGCCGCATCCAGCGCTGACGGAAGCAAAACACCAGCAGCAGTTTTTCCAGTTCCGTTTATCCGGGCAAACCATCGAAGTGACCAGCGAATACCTGTTCCGTCATAGCGATAACGAGCTCCTGCACTGGATGGTGGCGCTGGATGGTAAGCCGCTGGCAAGCGGTGAAGTGCCTCTGGATGTCGCTCCACAAGGTAAACAGTTGATTGAACTGCCTGAACTACCGCAGCCGGAGAGCGCCGGGCAACTCTGGCTCACAGTACGCGTAGTGCAACCGAACGCGACCGCATGGTCAGAAGCCGGGCACATCAGCGCCTGGCAGCAGTGGCGTCTGGCGGAAAACCTCAGTGTGACGCTCCCCGCCGCGTCCCACGCCATCCCGCATCTGACCACCAGCGAAATGGATTTTTGCATCGAGCTGGGTAATAAGCGTTGGCAATTTAACCGCCAGTCAGGCTTTCTTTCACAGATGTGGATTGGCGATAAAAAACAACTGCTGACGCCGCTGCGCGATCAGTTCACCCGTGCACCGCTGGATAACGACATTGGCGTAAGTGAAGCGACCCGCATTGACCCTAACGCCTGGGTCGAACGCTGGAAGGCGGCGGGCCATTACCAGGCCGAAGCAGCGTTGTTGCAGTGCACGGCAGATACACTTGCTGATGCGGTGCTGATTACGACCGCTCACGCGTGGCAGCATCAGGGGAAAACCTTATTTATCAGCCGGAAAACCTACCGGATTGATGGTAGTGGTCAAATGGCGATTACCGTTGATGTTGAAGTGGCGAGCGATACACCGCATCCGGCGCGGATTGGCCTGAACTGCCAGCTGGCGCAGGTAGCAGAGCGGGTAAACTGGCTCGGATTAGGGCCGCAAGAAAACTATCCCGACCGCCTTACTGCCGCCTGTTTTGACCGCTGGGATCTGCCATTGTCAGACATGTATACCCCGTACGTCTTCCCGAGCGAAAACGGTCTGCGCTGCGGGACGCGCGAATTGAATTATGGCCCACACCAGTGGCGCGGCGACTTCCAGTTCAACATCAGCCGCTACAGTCAACAGCAACTGATGGAAACCAGCCATCGCCATCTGCTGCACGCGGAAGAAGGCACATGGCTGAATATCGACGGTTTCCATATGGGGATTGGTGGCGACGACTCCTGGAGCCCGTCAGTATCGGCGGAATTCCAGCTGAGCGCCGGTCGCTACCATTACCAGTTGGTCTGGTGTCAAAAATAATAATAACCGGGCAGGCCATGTCTGCCCGTATTTCGCGTAAGGAAATCCATTATGTACTATTTAAAAAACACAAACTTTTGGATGTTCGGTTTATTCTTTTTCTTTTACTTTTTTATCATGGGAGCCTACTTCCCGTTTTTCCCGATTTGGCTACATGACATCAACCATATCAGCAAAAGTGATACGGGTATTATTTTTGCCGCTATTTCTCTGTTCTCGCTATTATTCCAACCGCTGTTTGGTCTGCTTTCTGACAAACTCGGGCTGCGCAAATACCTGCTGTGGATTATTACCGGCATGTTAGTGATGTTTGCGCCGTTCTTTATTTTTATCTTCGGGCCACTGTTACAATACAACATTTTAGTAGGATCGATTGTTGGTGGTATTTATCTAGGCTTTTGTTTTAACGCCGGTGCGCCAGCAGTAGAGGCATTTATTGAGAAAGTCAGCCGTCGCAGTAATTTCGAATTTGGTCGCGCGCGGATGTTTGGCTGTGTTGGCTGGGCGCTGTGTGCCTCGATTGTCGGCATCATGTTCACCATCAATAATCAGTTTGTTTTCTGGCTGGGCTCTGGCTGTGCACTCATCCTCGCCGTTTTACTCTTTTTCGCCAAAACGGATGCGCCCTCTTCTGCCACGGTTGCCAATGCGGTAGGTGCCAACCATTCGGCATTTAGCCTTAAGCTGGCACTGGAACTGTTCAGACAGCCAAAACTGTGGTTTTTGTCACTGTATGTTATTGGCGTTTCCTGCACCTACGATGTTTTTGACCAACAGTTTGCTAATTTCTTTACTTCGTTCTTTGCTACCGGTGAACAGGGTACGCGGGTATTTGGCTACGTAACGACAATGGGCGAATTACTTAACGCCTCGATTATGTTCTTTGCGCCACTGATCATTAATCGCATCGGTGGGAAAAACGCCCTGCTGCTGGCTGGCACTATTATGTCTGTACGTATTATTGGCTCATCGTTCGCCACCTCAGCGCTGGAAGTGGTTATTCTGAAAACGCTGCATATGTTTGAAGTACCGTTCCTGCTGGTGGGCTGCTTTAAATATATTACCAGCCAGTTTGAAGTGCGTTTTTCAGCGACGATTTATCTGGTCTGTTTCTGCTTCTTTAAGCAACTGGCGATGATTTTTATGTCTGTACTGGCGGGCAATATGTATGAAAGCATCGGTTTCCAGGGCGCTTATCTGGTGCTGGGTCTGGTGGCGCTGGGCTTCACCTTAATTTCCGTGTTCACGCTTAGCGGCCCCGGCCCGCTTTCCCTGCTGCGTCGTCAGGTGAATGAAGTCGCTTAAGCAATCAATGTCGGATGCGGCGCGAGCGCCTTATCCGACCAACATATCATAACGGAGTGATCGCATTGAACATGCCAATGACCGAAAGAATAAGAGCAGGCAAGCTATTTACCGATATGTGCGAAGGCTTACCGGAAAAAAGACTTCGTGGGAAAACGTTAATGTATGAGTTTAATCACTCGCATCCATCAGAAGTTGAAAAAAGAGAAAGCCTGATTAAAGAAATGTTTGCCACGGTAGGGGAAAACGCCTGGGTAGAACCGCCTGTCTATTTCTCTTACGGTTCCAACATCCATATAGGCCGCAATTTTTATGCAAATTTCAATTTAACCATTGTCGATGACTACACGGTAACAATCGGTGATAACGTACTGATTGCACCCAACGTTACTCTTTCCGTTACGGGACACCCTGTACACCATGAATTGAGAAAAAACGGCGAGATGTACTCTTTTCCGATAACGATTGGCAATAACGTCTGGATCGGAAGTCATGTGGTTATTAATCCAGGCGTCACCATCGGGGATAATTCTGTTATTGGCGCGGGTAGTATCGTCACAAAAGACATTCCACCAAACGTCGTGGCGGCTGGCGTTCCTTGTCGGGTTATTCGCGAAATAAACGACCGGGATAAGCACTATTATTTCAAAGATTATAAAGTTGAATCGTCAGTTTAAATTATAAAAATTGCCTGATACGCTGCGCTTATCAGGCCTACAAGTTCAGCGATCTACATTAGCCGCATCCGGCG

# Searching for information in a DNA sequence

You will do your own search for information contained in the DNA sequence of page 5:

* First, you search for Open Reading Frames (ORFs).
* Next, using an internet-based search algorithm, you identify the functions of the proteins that the ORFs encode.
* Finally, you search for sequences involved in regulation of gene transcription.

# Assignment-1: Search for genes in the DNA sequence

To find genes, the start and stop codons of the protein-encoding part of a gene need to be found within a reading frame. The start codon for almost all proteins is the triplet “ATG”, which translates into the amino acid methionine. The stop codon is either “TAA”, “TGA” or “TAG”. Stop codons do not encode for an amino acid but are recognized by release factor R, which terminates production of a polypeptide by the ribosome.

Your search for the protein-encoding part of a gene is now simplified to finding a start codon (ATG for methionine), which is followed by a string of triplets and a stop codon. The resulting DNA sequence is called an Open Reading Frame (ORF) and can be in any one of the six reading frames. In this course you will search for ORFs with more than 200 codons.

## Search for Open Reading Frames

[Click here](https://biochemistry-blast.containers.wur.nl/SMS/orf_find.html) to open the Sequence Manipulation Suite, which contains Bioinformatics tools for DNA and protein analysis.

* Select the entire DNA sequence on page 5 file and copy it by using the ‘ctrl-c’ command. Go to the program ‘ORF finder’ that you opened a moment ago. Remove the sample sequence from the input window using the button ‘clear’ and paste the input DNA sequence using ‘ctrl-v’ in the ORF finder.
* Now let the program search in all reading frames for ORFs that start with ATG (start codon) and are at least 200 codons long (see the screenshot of the ORF finder on the next page).

Screenshot of the ORF finder:

### Graphical user interface, text, application Description automatically generated

* You will first examine the direct strand of the nucleotide sequence by clicking the submit button. **Copy the resulting DNA and amino acid sequences and paste them in the box below:**

### ORFs found in the direct strand

(Keep the formatting of the original text, by using the “keep text only” option)

ORF Finder results

**Results for 6400 residue sequence "Untitled" starting "GCGCAAAACC"**

>ORF number 1 in reading frame 1 on the direct strand extends from base 1270 to base 4344.

ATGACCATGATTACGGATTCACTGGCCGTCGTTTTACAACGTCGTGACTGGGAAAACCCT

GGCGTTACCCAACTTAATCGCCTTGCAGCACATCCCCCTTTCGCCAGCTGGCGTAATAGC

GAAGAGGCCCGCACCGATCGCCCTTCCCAACAGTTGCGCAGCCTGAATGGCGAATGGCGC

TTTGCCTGGTTTCCGGCACCAGAAGCGGTGCCGGAAAGCTGGCTGGAGTGCGATCTTCCT

GAGGCCGATACTGTCGTCGTCCCCTCAAACTGGCAGATGCACGGTTACGATGCGCCCATC

TACACCAACGTGACCTATCCCATTACGGTCAATCCGCCGTTTGTTCCCACGGAGAATCCG

ACGGGTTGTTACTCGCTCACATTTAATGTTGATGAAAGCTGGCTACAGGAAGGCCAGACG

CGAATTATTTTTGATGGCGTTAACTCGGCGTTTCATCTGTGGTGCAACGGGCGCTGGGTC

GGTTACGGCCAGGACAGTCGTTTGCCGTCTGAATTTGACCTGAGCGCATTTTTACGCGCC

GGAGAAAACCGCCTCGCGGTGATGGTGCTGCGCTGGAGTGACGGCAGTTATCTGGAAGAT

CAGGATATGTGGCGGATGAGCGGCATTTTCCGTGACGTCTCGTTGCTGCATAAACCGACT

ACACAAATCAGCGATTTCCATGTTGCCACTCGCTTTAATGATGATTTCAGCCGCGCTGTA

CTGGAGGCTGAAGTTCAGATGTGCGGCGAGTTGCGTGACTACCTACGGGTAACAGTTTCT

TTATGGCAGGGTGAAACGCAGGTCGCCAGCGGCACCGCGCCTTTCGGCGGTGAAATTATC

GATGAGCGTGGTGGTTATGCCGATCGCGTCACACTACGTCTGAACGTCGAAAACCCGAAA

CTGTGGAGCGCCGAAATCCCGAATCTCTATCGTGCGGTGGTTGAACTGCACACCGCCGAC

GGCACGCTGATTGAAGCAGAAGCCTGCGATGTCGGTTTCCGCGAGGTGCGGATTGAAAAT

GGTCTGCTGCTGCTGAACGGCAAGCCGTTGCTGATTCGAGGCGTTAACCGTCACGAGCAT

CATCCTCTGCATGGTCAGGTCATGGATGAGCAGACGATGGTGCAGGATATCCTGCTGATG

AAGCAGAACAACTTTAACGCCGTGCGCTGTTCGCATTATCCGAACCATCCGCTGTGGTAC

ACGCTGTGCGACCGCTACGGCCTGTATGTGGTGGATGAAGCCAATATTGAAACCCACGGC

ATGGTGCCAATGAATCGTCTGACCGATGATCCGCGCTGGCTACCGGCGATGAGCGAACGC

GTAACGCGAATGGTGCAGCGCGATCGTAATCACCCGAGTGTGATCATCTGGTCGCTGGGG

AATGAATCAGGCCACGGCGCTAATCACGACGCGCTGTATCGCTGGATCAAATCTGTCGAT

CCTTCCCGCCCGGTGCAGTATGAAGGCGGCGGAGCCGACACCACGGCCACCGATATTATT

TGCCCGATGTACGCGCGCGTGGATGAAGACCAGCCCTTCCCGGCTGTGCCGAAATGGTCC

ATCAAAAAATGGCTTTCGCTACCTGGAGAGACGCGCCCGCTGATCCTTTGCGAATACGCC

CACGCGATGGGTAACAGTCTTGGCGGTTTCGCTAAATACTGGCAGGCGTTTCGTCAGTAT

CCCCGTTTACAGGGCGGCTTCGTCTGGGACTGGGTGGATCAGTCGCTGATTAAATATGAT

GAAAACGGCAACCCGTGGTCGGCTTACGGCGGTGATTTTGGCGATACGCCGAACGATCGC

CAGTTCTGTATGAACGGTCTGGTCTTTGCCGACCGCACGCCGCATCCAGCGCTGACGGAA

GCAAAACACCAGCAGCAGTTTTTCCAGTTCCGTTTATCCGGGCAAACCATCGAAGTGACC

AGCGAATACCTGTTCCGTCATAGCGATAACGAGCTCCTGCACTGGATGGTGGCGCTGGAT

GGTAAGCCGCTGGCAAGCGGTGAAGTGCCTCTGGATGTCGCTCCACAAGGTAAACAGTTG

ATTGAACTGCCTGAACTACCGCAGCCGGAGAGCGCCGGGCAACTCTGGCTCACAGTACGC

GTAGTGCAACCGAACGCGACCGCATGGTCAGAAGCCGGGCACATCAGCGCCTGGCAGCAG

TGGCGTCTGGCGGAAAACCTCAGTGTGACGCTCCCCGCCGCGTCCCACGCCATCCCGCAT

CTGACCACCAGCGAAATGGATTTTTGCATCGAGCTGGGTAATAAGCGTTGGCAATTTAAC

CGCCAGTCAGGCTTTCTTTCACAGATGTGGATTGGCGATAAAAAACAACTGCTGACGCCG

CTGCGCGATCAGTTCACCCGTGCACCGCTGGATAACGACATTGGCGTAAGTGAAGCGACC

CGCATTGACCCTAACGCCTGGGTCGAACGCTGGAAGGCGGCGGGCCATTACCAGGCCGAA

GCAGCGTTGTTGCAGTGCACGGCAGATACACTTGCTGATGCGGTGCTGATTACGACCGCT

CACGCGTGGCAGCATCAGGGGAAAACCTTATTTATCAGCCGGAAAACCTACCGGATTGAT

GGTAGTGGTCAAATGGCGATTACCGTTGATGTTGAAGTGGCGAGCGATACACCGCATCCG

GCGCGGATTGGCCTGAACTGCCAGCTGGCGCAGGTAGCAGAGCGGGTAAACTGGCTCGGA

TTAGGGCCGCAAGAAAACTATCCCGACCGCCTTACTGCCGCCTGTTTTGACCGCTGGGAT

CTGCCATTGTCAGACATGTATACCCCGTACGTCTTCCCGAGCGAAAACGGTCTGCGCTGC

GGGACGCGCGAATTGAATTATGGCCCACACCAGTGGCGCGGCGACTTCCAGTTCAACATC

AGCCGCTACAGTCAACAGCAACTGATGGAAACCAGCCATCGCCATCTGCTGCACGCGGAA

GAAGGCACATGGCTGAATATCGACGGTTTCCATATGGGGATTGGTGGCGACGACTCCTGG

AGCCCGTCAGTATCGGCGGAATTCCAGCTGAGCGCCGGTCGCTACCATTACCAGTTGGTC

TGGTGTCAAAAATAA

>Translation of ORF number 1 in reading frame 1 on the direct strand.

MTMITDSLAVVLQRRDWENPGVTQLNRLAAHPPFASWRNSEEARTDRPSQQLRSLNGEWR

FAWFPAPEAVPESWLECDLPEADTVVVPSNWQMHGYDAPIYTNVTYPITVNPPFVPTENP

TGCYSLTFNVDESWLQEGQTRIIFDGVNSAFHLWCNGRWVGYGQDSRLPSEFDLSAFLRA

GENRLAVMVLRWSDGSYLEDQDMWRMSGIFRDVSLLHKPTTQISDFHVATRFNDDFSRAV

LEAEVQMCGELRDYLRVTVSLWQGETQVASGTAPFGGEIIDERGGYADRVTLRLNVENPK

LWSAEIPNLYRAVVELHTADGTLIEAEACDVGFREVRIENGLLLLNGKPLLIRGVNRHEH

HPLHGQVMDEQTMVQDILLMKQNNFNAVRCSHYPNHPLWYTLCDRYGLYVVDEANIETHG

MVPMNRLTDDPRWLPAMSERVTRMVQRDRNHPSVIIWSLGNESGHGANHDALYRWIKSVD

PSRPVQYEGGGADTTATDIICPMYARVDEDQPFPAVPKWSIKKWLSLPGETRPLILCEYA

HAMGNSLGGFAKYWQAFRQYPRLQGGFVWDWVDQSLIKYDENGNPWSAYGGDFGDTPNDR

QFCMNGLVFADRTPHPALTEAKHQQQFFQFRLSGQTIEVTSEYLFRHSDNELLHWMVALD

GKPLASGEVPLDVAPQGKQLIELPELPQPESAGQLWLTVRVVQPNATAWSEAGHISAWQQ

WRLAENLSVTLPAASHAIPHLTTSEMDFCIELGNKRWQFNRQSGFLSQMWIGDKKQLLTP

LRDQFTRAPLDNDIGVSEATRIDPNAWVERWKAAGHYQAEAALLQCTADTLADAVLITTA

HAWQHQGKTLFISRKTYRIDGSGQMAITVDVEVASDTPHPARIGLNCQLAQVAERVNWLG

LGPQENYPDRLTAACFDRWDLPLSDMYTPYVFPSENGLRCGTRELNYGPHQWRGDFQFNI

SRYSQQQLMETSHRHLLHAEEGTWLNIDGFHMGIGGDDSWSPSVSAEFQLSAGRYHYQLV

WCQK\*

>ORF number 2 in reading frame 1 on the direct strand extends from base 4396 to base 5649.

ATGTACTATTTAAAAAACACAAACTTTTGGATGTTCGGTTTATTCTTTTTCTTTTACTTT

TTTATCATGGGAGCCTACTTCCCGTTTTTCCCGATTTGGCTACATGACATCAACCATATC

AGCAAAAGTGATACGGGTATTATTTTTGCCGCTATTTCTCTGTTCTCGCTATTATTCCAA

CCGCTGTTTGGTCTGCTTTCTGACAAACTCGGGCTGCGCAAATACCTGCTGTGGATTATT

ACCGGCATGTTAGTGATGTTTGCGCCGTTCTTTATTTTTATCTTCGGGCCACTGTTACAA

TACAACATTTTAGTAGGATCGATTGTTGGTGGTATTTATCTAGGCTTTTGTTTTAACGCC

GGTGCGCCAGCAGTAGAGGCATTTATTGAGAAAGTCAGCCGTCGCAGTAATTTCGAATTT

GGTCGCGCGCGGATGTTTGGCTGTGTTGGCTGGGCGCTGTGTGCCTCGATTGTCGGCATC

ATGTTCACCATCAATAATCAGTTTGTTTTCTGGCTGGGCTCTGGCTGTGCACTCATCCTC

GCCGTTTTACTCTTTTTCGCCAAAACGGATGCGCCCTCTTCTGCCACGGTTGCCAATGCG

GTAGGTGCCAACCATTCGGCATTTAGCCTTAAGCTGGCACTGGAACTGTTCAGACAGCCA

AAACTGTGGTTTTTGTCACTGTATGTTATTGGCGTTTCCTGCACCTACGATGTTTTTGAC

CAACAGTTTGCTAATTTCTTTACTTCGTTCTTTGCTACCGGTGAACAGGGTACGCGGGTA

TTTGGCTACGTAACGACAATGGGCGAATTACTTAACGCCTCGATTATGTTCTTTGCGCCA

CTGATCATTAATCGCATCGGTGGGAAAAACGCCCTGCTGCTGGCTGGCACTATTATGTCT

GTACGTATTATTGGCTCATCGTTCGCCACCTCAGCGCTGGAAGTGGTTATTCTGAAAACG

CTGCATATGTTTGAAGTACCGTTCCTGCTGGTGGGCTGCTTTAAATATATTACCAGCCAG

TTTGAAGTGCGTTTTTCAGCGACGATTTATCTGGTCTGTTTCTGCTTCTTTAAGCAACTG

GCGATGATTTTTATGTCTGTACTGGCGGGCAATATGTATGAAAGCATCGGTTTCCAGGGC

GCTTATCTGGTGCTGGGTCTGGTGGCGCTGGGCTTCACCTTAATTTCCGTGTTCACGCTT

AGCGGCCCCGGCCCGCTTTCCCTGCTGCGTCGTCAGGTGAATGAAGTCGCTTAA

>Translation of ORF number 2 in reading frame 1 on the direct strand.

MYYLKNTNFWMFGLFFFFYFFIMGAYFPFFPIWLHDINHISKSDTGIIFAAISLFSLLFQ

PLFGLLSDKLGLRKYLLWIITGMLVMFAPFFIFIFGPLLQYNILVGSIVGGIYLGFCFNA

GAPAVEAFIEKVSRRSNFEFGRARMFGCVGWALCASIVGIMFTINNQFVFWLGSGCALIL

AVLLFFAKTDAPSSATVANAVGANHSAFSLKLALELFRQPKLWFLSLYVIGVSCTYDVFD

QQFANFFTSFFATGEQGTRVFGYVTTMGELLNASIMFFAPLIINRIGGKNALLLAGTIMS

VRIIGSSFATSALEVVILKTLHMFEVPFLLVGCFKYITSQFEVRFSATIYLVCFCFFKQL

AMIFMSVLAGNMYESIGFQGAYLVLGLVALGFTLISVFTLSGPGPLSLLRRQVNEVA\*

>ORF number 1 in reading frame 2 on the direct strand extends from base 65 to base 1147.

ATGAAACCAGTAACGTTATACGATGTCGCAGAGTATGCCGGTGTCTCTTATCAGACCGTT

TCCCGCGTGGTGAACCAGGCCAGCCACGTTTCTGCGAAAACGCGGGAAAAAGTGGAAGCG

GCGATGGCGGAGCTGAATTACATTCCCAACCGCGTGGCACAACAACTGGCGGGCAAACAG

TCGTTGCTGATTGGCGTTGCCACCTCCAGTCTGGCCCTGCACGCGCCGTCGCAAATTGTC

GCGGCGATTAAATCTCGCGCCGATCAACTGGGTGCCAGCGTGGTGGTGTCGATGGTAGAA

CGAAGCGGCGTCGAAGCCTGTAAAGCGGCGGTGCACAATCTTCTCGCGCAACGCGTCAGT

GGGCTGATCATTAACTATCCGCTGGATGACCAGGATGCCATTGCTGTGGAAGCTGCCTGC

ACTAATGTTCCGGCGTTATTTCTTGATGTCTCTGACCAGACACCCATCAACAGTATTATT

TTCTCCCATGAAGACGGTACGCGACTGGGCGTGGAGCATCTGGTCGCATTGGGTCACCAG

CAAATCGCGCTGTTAGCGGGCCCATTAAGTTCTGTCTCGGCGCGTCTGCGTCTGGCTGGC

TGGCATAAATATCTCACTCGCAATCAAATTCAGCCGATAGCGGAACGGGAAGGCGACTGG

AGTGCCATGTCCGGTTTTCAACAAACCATGCAAATGCTGAATGAGGGCATCGTTCCCACT

GCGATGCTGGTTGCCAACGATCAGATGGCGCTGGGCGCAATGCGCGCCATTACCGAGTCC

GGGCTGCGCGTTGGTGCGGATATCTCGGTAGTGGGATACGACGATACCGAAGACAGCTCA

TGTTATATCCCGCCGTTAACCACCATCAAACAGGATTTTCGCCTGCTGGGGCAAACCAGC

GTGGACCGCTTGCTGCAACTCTCTCAGGGCCAGGCGGTGAAGGGCAATCAGCTGTTGCCC

GTCTCACTGGTGAAAAGAAAAACCACCCTGGCGCCCAATACGCAAACCGCCTCTCCCCGC

GCGTTGGCCGATTCATTAATGCAGCTGGCACGACAGGTTTCCCGACTGGAAAGCGGGCAG

TGA

>Translation of ORF number 1 in reading frame 2 on the direct strand.

MKPVTLYDVAEYAGVSYQTVSRVVNQASHVSAKTREKVEAAMAELNYIPNRVAQQLAGKQ

SLLIGVATSSLALHAPSQIVAAIKSRADQLGASVVVSMVERSGVEACKAAVHNLLAQRVS

GLIINYPLDDQDAIAVEAACTNVPALFLDVSDQTPINSIIFSHEDGTRLGVEHLVALGHQ

QIALLAGPLSSVSARLRLAGWHKYLTRNQIQPIAEREGDWSAMSGFQQTMQMLNEGIVPT

AMLVANDQMALGAMRAITESGLRVGADISVVGYDDTEDSSCYIPPLTTIKQDFRLLGQTS

VDRLLQLSQGQAVKGNQLLPVSLVKRKTTLAPNTQTASPRALADSLMQLARQVSRLESGQ

\*

>ORF number 1 in reading frame 3 on the direct strand extends from base 5664 to base 6326.

ATGCGGCGCGAGCGCCTTATCCGACCAACATATCATAACGGAGTGATCGCATTGAACATG

CCAATGACCGAAAGAATAAGAGCAGGCAAGCTATTTACCGATATGTGCGAAGGCTTACCG

GAAAAAAGACTTCGTGGGAAAACGTTAATGTATGAGTTTAATCACTCGCATCCATCAGAA

GTTGAAAAAAGAGAAAGCCTGATTAAAGAAATGTTTGCCACGGTAGGGGAAAACGCCTGG

GTAGAACCGCCTGTCTATTTCTCTTACGGTTCCAACATCCATATAGGCCGCAATTTTTAT

GCAAATTTCAATTTAACCATTGTCGATGACTACACGGTAACAATCGGTGATAACGTACTG

ATTGCACCCAACGTTACTCTTTCCGTTACGGGACACCCTGTACACCATGAATTGAGAAAA

AACGGCGAGATGTACTCTTTTCCGATAACGATTGGCAATAACGTCTGGATCGGAAGTCAT

GTGGTTATTAATCCAGGCGTCACCATCGGGGATAATTCTGTTATTGGCGCGGGTAGTATC

GTCACAAAAGACATTCCACCAAACGTCGTGGCGGCTGGCGTTCCTTGTCGGGTTATTCGC

GAAATAAACGACCGGGATAAGCACTATTATTTCAAAGATTATAAAGTTGAATCGTCAGTT

TAA

>Translation of ORF number 1 in reading frame 3 on the direct strand.

MRRERLIRPTYHNGVIALNMPMTERIRAGKLFTDMCEGLPEKRLRGKTLMYEFNHSHPSE

VEKRESLIKEMFATVGENAWVEPPVYFSYGSNIHIGRNFYANFNLTIVDDYTVTIGDNVL

IAPNVTLSVTGHPVHHELRKNGEMYSFPITIGNNVWIGSHVVINPGVTIGDNSVIGAGSI

VTKDIPPNVVAAGVPCRVIREINDRDKHYYFKDYKVESSV\*

* Now, look for ORFs in the reverse strand and **copy the results and paste them in the box below:**

### ORFs found in the reverse strand

(Keep the formatting of the original text, by using the “keep text only” option)

ORF Finder results

Results for 6400 residue sequence "Untitled" starting "GCGCAAAACC"

No ORFs were found in reading frame 1.

No ORFs were found in reading frame 2.

No ORFs were found in reading frame 3.

## Question-1: How many ORFs did you identify in each of the 6 reading frames?

Fill out the table below

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ORF number | Reading frame | First base number | Last base number | Total number of bases | Number of amino acids |
| 1 | 1 | 1270 | 4344 |  |  |
| 2 | 1 | 4396 | 5649 |  |  |
| 1 | 2 | 65 | 1147 |  |  |
| 1 | 3 | 5664 | 6326 |  |  |
|  |  |  |  |  |  |

## Question-2: Highlight with different colours the ORFS in the DNA sequence on page 5.

Note that the output of the ORF finder is sorted by reading frame number.

# Assignment-2: Identification of DNA and/or protein sequences

Expression of a gene into protein starts with the transcription of the corresponding nucleotide sequence into an mRNA molecule by RNA-polymerase. This enzyme reads the template strand of a DNA molecule in the 3’-->5’ direction and synthesizes the corresponding mRNA molecule in the 5’-->3’ direction. The nucleotide sequence of the resulting mRNA molecule is identical to the coding DNA strand. Note that RNA contains ribonucleotides instead of deoxyribonucleotides and uracil (U) instead of thymidine (T). Next, a ribosome can bind to mRNA and translate it from the 5’-->3’ direction into a polypeptide chain (or protein).

The output of the ORF finder you used includes the amino acid sequences of the encoded proteins. Now you can identify the function of these proteins based on their amino acid sequences.

DNA and protein sequences from many organisms are stored in databases. When the function of a certain protein has been characterised, this information is included. A way of searching a database to identify a protein is by using so-called BLAST algorithms. A BLAST algorithm searches for small local alignments (matches) and reports only those that score above a threshold. When you submit your amino acid sequence of interest (“query” sequence), the BLAST algorithm retrieves from the database identical or similar amino acid sequences, which are called “hits”. So, BLAST can be used to identify your protein of interest.

## Use BLAST to identify proteins

* You will now run your own BLAST search. Select and copy the amino acid sequence of the first ORF you have found. [Clic k here](https://biochemistry-blast.containers.wur.nl/) to open the BLAST website. Paste the amino acid sequence into the window after you removed the sample sequence. The ‘output’ (i.e., maximum number of hits) is limited to 5 as a standard value. Press the button marked BLAST.
* Inspect the results obtained. Your input sequence is always on top and marked as “Query” while the amino acid sequence that is retrieved from the BLAST database is the bottom one (“Hit”). Shown in between are the amino acids that are identical in Query and Hit (“alignment”). **Write down which protein is encoded by the first ORF in the box below:**

### The protein encoded by the first ORF is:

MTMITDSLAVVLQRRDWENPGVTQLNRLAAHPPFASWRNSEEARTDRPSQQLRSLNGEWR

* Repeat the BLAST by changing the ‘output’ from ‘5’ to ‘all’, look at the alignments on the bottom of the page (you will need to scroll down to see them), and note the differences between both BLAST searches. **Copy a BLAST result that contains hyphen (-) as well as (+) signs, and paste it in the box below:**

### A BLAST result that contains hyphen (-) and (+) signs:

(Keep the formatting of the original text, by using the “keep text only” option)

Query: QFCMNGLVFADRTPHPALTEAKHQQQFFQFRL---SGQTIEVTSEYLFRHSDNELLHWMV

Alignment: QFCMNGLVFADRTPHPAL EAKH QQFFQFRL + IEV SEYLFRHSDNE+L WM+

Hit: QFCMNGLVFADRTPHPALYEAKHVQQFFQFRLLPGEERRIEVQSEYLFRHSDNEILRWML

## Question-3: What do the hyphen ( - ) and ( + ) signs in the query, hit, or alignment sequences mean?

(Hint: compare the properties of amino acids.)

( - ) is a gap in the sequence alignment

( + ) means that the sequence in the alignment is in the same orientation as it is in the database

## Question-4: What is the identity of the proteins encoded in the other ORFs?

**Use BLAST to identify the other ORFs (use the standard limit of 5 in the output).** Fill out the table below

|  |  |
| --- | --- |
| Position in sequence | Protein identity |
| 1 |  |
| 2 |  |
| 3 |  |
| 4 |  |

# Assignment-3: Finding regulatory elements in a nucleotide sequence

ORFs are not the only parts of genes that can be recognized in a DNA sequence. DNA sequences that play a role in regulation of DNA transcription are usually located before the start of an ORF (molecular biologists call this ‘upstream’). For example, the nucleotide sequence of the promoter is about 100-200 nucleotides upstream. The promoter is the binding site for RNA-polymerase and is the starting point for transcription of DNA into mRNA. For example, the binding site for the ribosome is about 20-50 nucleotides upstream of the ORF. In mRNA this site is essential for binding of ribosomes to start translation.

Downstream of an ORF are nucleotide sequences that indicate that transcription should stop (‘terminator sequences’).

Regulatory DNA sequences may differ but are similar enough to recognize them as a ‘pattern’ or ‘consensus sequence’. For instance, a nucleotide sequence that resembles “TATA” is often found in the promoter.

## Locating promoter and operator sequences.

* You are first going to search for the sequence ‘TATA' in the sequence on page 5 using the “find” function in Word (Ctrl+f).

## Question-5: How many times and in which sequences do you find **TATA** in the direct strand? Where do you expect promoter sequences?

7, at the 5’ end

## Question-6: Identify the position of the promoter sequence directly in front of the LacZ gene.

Often TATA sequences are not completely conserved. For instance, the promoter of the LacZ gene has the sequence **CGTATGTTGTGTGGA** (which contains “TATG” instead of “TATA”).

**Use this sequence to find the promoter in the DNA sequence on page 5 and highlight it.**

## Question-7: There is a single promoter in front of LacZ. Your search showed that there is no promoter in front of LacY or LacA. Why is this?

There is no separate promoter for LacY and LacA because the three genes (LacZ, LacY, and LacA) in the lac operon are transcribed together as a single polycistronic mRNA from a single promoter located upstream of LacZ

## Question-8: On basis of the identity of ORFs 2, 3 and 4, which operon do you identify?

## Question-9: What is the position of the Lac operator sequence?

**TGTGTGGAATTGTGAGCGGATAACAATTTCACACA** is the Lac operator sequence. **Now use this sequence to find the Lac operator in the DNA sequence on page 5 and highlight it.**

## Question-10: What is the position of the binding site for catabolite activator protein?

Catabolite activator protein (CAP) is another key player in the regulation of the Lac operon. Upon activation of CAP by cAMP it binds to the DNA sequence **TAATGTGAGTTAGCTCACTCAT** and as a result transcription becomes more efficient. **Find and highlight the CAP binding site in the DNA sequence on page 5.**

# Assignment-4: do the Bioinformatics test on Brightspace

Performing the test is mandatory for completing the practical of week 2.