# ASSIGNMENT 2
# BIOINFORMATICS (V1.01)

## Specification

For this assignment, write a program that accesses and analyses protein sequences. The protein sequences that you will analyse come from GenBank®, which is the National Institutes of Health (USA) genetic sequence database[1]. A full list of DNA, RNA, and protein sequences can be found on their FTP site[2]. For this assignment, use the abridged (`protein_a.fa`) and complete (`protein_c.fa`) files provided on blackboard.

**1. Selection of database.** Create a user interface that helps navigate through the different proteins. At the start of the program, the user should be asked whether he or she would like to use the abridged (option `1`) or complete (option `2`) database. The selected database should then be loaded into the program. **(NB: Lecture 5 will describe key concepts that will be useful here)**

**2. Main menu.** The user should then be presented with a menu of options.

1. Overview of this database. List the total number of proteins in the database and display a table that associates each amino acid character to its proper name (e.g., '`A: Alanine`').
2. Search by item id. This id refers to the protein in the order that it appears in the database.
3. Search by gi id. This id is the first identification labelled in the description line.
4. Search by ref id. This id is the second identification labelled in the description line.
5. Search by keyword. Search through the series of words within the third identification label of all the proteins within the database for a keyword match. A match is made if the keyword is part of another word or alone. If a match is found, display the protein details to the users and ask whether this was the protein he or she was seeking. If multiple matches are found, display the options and allow the user to choose which one to select or none of the above.
6. Quit the program.

**3. Search hit menu.** When a match is made, another set of options should be displayed.

1. Description. Display the gi id, the ref id, and the name of the protein.
2. Protein sequence. Display the full sequence.
3. Protein statistics. Display the number of amino acids in the sequence for each amino acid.
4. Record protein to file. Record the selected protein to a file called '`selected_proteins.txt`'.
5. Return to main menu.

## Notes and Tips

Successful implementation of the abridged (`protein_a.fa`) database without a keyword search option will be awarded 80% of the total mark. The remaining 20% will be awarded to the full implementation. I suggest starting with `protein_a.fa` before moving on to the larger database. Begin by learning how to open, read, and close files. Then proceed to storing the contents of the database into your program's data structures.

## Submission Procedure

Submit only one file labelled '`proteins.cpp`' through the "Assignment 2 – Submission Portal" on blackboard (i.e., do not submit the `.fa` files). **The submission deadline is 4 PM on 21 November 2016**. Submissions are NOT possible after this deadline.

## References

[1] https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Homo_sapiens/108/
[2] ftp://ftp.ncbi.nlm.nih.gov/genomes/Homo_sapiens

### Example Program Output

```
Welcome to the Protein Database

Select an option from the menu below
1) Load the abridged protein data
2) Load the complete protein data
3) Quit database
>>2

Loading database...
Database loaded.

Select an option from the menu below
1) Overview of the database
2) Search by protein #
3) Search by gi #
4) Search by ref #
5) Search by keyword
6) Quit database
>>1

The proteins in the database are from GenBank(R)
Total number of proteins in the database: 110386
Amino acids are represented by the following characters:
A  alanine                P  proline
B  aspartate/asparagine   Q  glutamine
C  cystine                R  arginine
D  aspartate              S  serine
E  glutamate              T  threonine
F  phenylalanine          U  selenocysteine
G  glycine                V  valine
H  histidine              W  tryptophan
I  isoleucine             Y  tyrosine
K  lysine                 Z  glutamate/glutamine
L  leucine                X  any
M  methionine             *  translation stop
N  asparagine             -  gap of indeterminate length

Select an option from the menu below
1) Overview of the database
2) Search by protein #
3) Search by gi #
4) Search by ref #
5) Search by keyword
6) Quit database
>>2

Enter an item id
>>6329

Select and option from the menu below
1) Description of the protein
2) Protein sequence
3) Protein statistics
4) Record protein to file
5) Return to main menu
>>1

Description of the protein:
item id: 6329
gi id:   752992854
ref id:  NP_001291652.1
name:     pleckstrin homology domain-containing family O member 1 isoform b [Homo
sapiens]
```

```
Select and option from the menu below
1) Description of the protein
2) Protein sequence
3) Protein statistics
4) Record protein to file
5) Return to main menu
>>2

Protein sequence:
MAVASTSTSDGMLTLDLIQEEDPSPEEPTSCAESFRVDLDKSVAQLAGSRRRADSDRIQPSADRASSLSRPWEKTDKGATYTPQAPKKL
TPTEKGRCASLEEILSQRDAASARTLQLRAEEPPTPALPNPGQLSRIQDLVARKLEETQELLAEVQGLGDGKRKAKDPPRSPPDSESEQ
LLLETERLLGEASSNWSQAKRVLQEVRELRDLYRQMDLQTPDSHLRQTTPHSQYRKSLM

Select and option from the menu below
1) Description of the protein
2) Protein sequence
3) Protein statistics
4) Record protein to file
5) Return to main menu
>>3

Report on the protein statistics:

Total number of amino acids: 237
A  21P 19
B  0 Q 17
C  2 R 22
D  17S 26
E  22T 16
F  1 U 0
G  9 V 7
H  2 W 2
I  4 Y 3
K  12Z 0
L  29X 0
M  4 * 0
N  2 - 0

Select and option from the menu below
1) Description of the protein
2) Protein sequence
3) Protein statistics
4) Record protein to file
5) Return to main menu
>>4

The protein was written to file.

Select and option from the menu below
1) Description of the protein
2) Protein sequence
3) Protein statistics
4) Record protein to file
5) Return to main menu
>>5

Select an option from the menu below
1) Overview of the database
2) Search by protein #
3) Search by gi #
4) Search by ref #
5) Search by keyword
6) Quit database
>>5

Enter a keyword
>>hemoglobin
```

```
Number of matches founds: 15
1) item id: 59405, gi id: 4504349, ref id: NP_000509.1
    hemoglobin subunit beta [Homo sapiens]
2) item id: 59406, gi id: 4504351, ref id: NP_000510.1
    hemoglobin subunit delta [Homo sapiens]
3) item id: 59407, gi id: 28302131, ref id: NP_000550.2
    hemoglobin subunit gamma-1 [Homo sapiens]
4) item id: 59408, gi id: 6715607, ref id: NP_000175.1
    hemoglobin subunit gamma-2 [Homo sapiens]
5) item id: 59409, gi id: 4885393, ref id: NP_005321.1
    hemoglobin subunit epsilon [Homo sapiens]
6) item id: 79336, gi id: 530407992, ref id: XP_005255344.1
    PREDICTED: hemoglobin subunit zeta isoform X1 [Homo sapiens]
7) item id: 79337, gi id: 4885397, ref id: NP_005323.1
    hemoglobin subunit zeta [Homo sapiens]
8) item id: 79338, gi id: 530407994, ref id: XP_005255345.1
    PREDICTED: hemoglobin subunit zeta isoform X2 [Homo sapiens]
9) item id: 79339, gi id: 51510893, ref id: NP_001003938.1
    hemoglobin subunit mu [Homo sapiens]
10) item id: 79340, gi id: 4504345, ref id: NP_000508.1
    hemoglobin subunit alpha [Homo sapiens]
11) item id: 79341, gi id: 4504347, ref id: NP_000549.1
    hemoglobin subunit alpha [Homo sapiens]
12) item id: 79342, gi id: 4885395, ref id: NP_005322.1
    hemoglobin subunit theta-1 [Homo sapiens]
13) item id: 81647, gi id: 7706180, ref id: NP_057717.1
    alpha-hemoglobin-stabilizing protein [Homo sapiens]
14) item id: 81648, gi id: 970841892, ref id: NP_001305150.1
    alpha-hemoglobin-stabilizing protein [Homo sapiens]
15) item id: 81649, gi id: 970841894, ref id: NP_001305151.1
    alpha-hemoglobin-stabilizing protein [Homo sapiens]

Select one of the matches
>>4

59408Select and option from the menu below
1) Description of the protein
2) Protein sequence
3) Protein statistics
4) Record protein to file
5) Return to main menu
>>1

Description of the protein:
item id: 59409
gi id:   4885393
ref id:  NP_005321.1
name:     hemoglobin subunit epsilon [Homo sapiens]

Select and option from the menu below
1) Description of the protein
2) Protein sequence
3) Protein statistics
4) Record protein to file
5) Return to main menu
>>2

Protein sequence:
MVHFTAEEKAAVTSLWSKMNVEEAGGEALGRLLVVYPWTQRFFDSFGNLSSPSAILGNPKVKAHGKKVLTSFGDAIKNMDNLKPAFAKL
SELHCDKLHVDPENFKLLGNVMVIILATHFGKEFTPEVQAAWQKLVSAVAIALAHKYH
```

```
Select and option from the menu below
1) Description of the protein
2) Protein sequence
3) Protein statistics
4) Record protein to file
5) Return to main menu
>>3

Report on the protein statistics:

Total number of amino acids: 147
A  17P 6
B  0 Q 3
C  1 R 2
D  5 S 9
E  9 T 6
F  9 U 0
G  9 V 13
H  7 W 3
I  5 Y 2
K  14Z 0
L  16X 0
M  4 * 0
N  7 - 0

Select and option from the menu below
1) Description of the protein
2) Protein sequence
3) Protein statistics
4) Record protein to file
5) Return to main menu
>>4

The protein was written to file.

Select and option from the menu below
1) Description of the protein
2) Protein sequence
3) Protein statistics
4) Record protein to file
5) Return to main menu
>>5

Select an option from the menu below
1) Overview of the database
2) Search by protein #
3) Search by gi #
4) Search by ref #
5) Search by keyword
6) Quit database
>>6

Exiting the database
```