Assignment 2
DNA SEQUENCE DATABASE

Deadline: 4 PM on 8 November 2019

Submission procedure: Submit only one file labelled '`a2.cpp`' through the TurnItIn portal on blackboard.

Specification

Define an object type that a programmer can store and analyse DNA sequences. The object type should have basic database functions, such as the loading and saving of sequences, analysis of sequences, and searching within sequences. You must use object-oriented programming. Define at least one `class` type for the database, but 2 or more classes may allow for easier implementation. Name your DNA sequence database object type `DNA_DB`. For example, to create a database object, you would write `DNA_DB dna_db;` in your `main()` function.

Implement an opening menu:

```
DNA Sequence Database Software

Specify the name of DNA sequence file names you would like to load. For
multiple files, add a ',' between each file name.
>
```

The `>` specifies to the user that an input is required. The user will specify the name of DNA sequences written in the FASTA format (`.fa` files). If all files are valid, then load the DNA sequences into your DNA sequence database object, otherwise, ask the user to specify the names again. Multiple file names can be entered by separating the names using a comma (`,`).

Once the DNA sequences are loaded, display the following menu:

```
Select one of the following options
(S) Summary statistics of the DNA database
(1) Analyse DNA_sequence_1
(2) Analyse DNA_sequence_2
(3) Analyse DNA_sequence_3
(4) Analyse DNA_sequence_4
(Q) Quit
>
```

Option S should display the number of sequences loaded into the database. For each sequence, display the sequence name and identifiers and the number of base pairs.

Option Q should quit the program.

Numeric options, (e.g., 1-4 in this example), should list the name of each DNA sequence loaded into the database (e.g., file name). Selection of a number should bring up the subsequent menu.

Once a numeric option is selected, display the following menu:

```
Select one of the following options
(H) Help
(S) Summary statistics of the DNA sequence
(1) Analyse gap region
(2) Analyse coded region
(3) Analyse base pair range
(4) Find DNA sequence by manual input
(5) Find DNA sequence by file input
(R) Return to the previous menu
(Q) Quit
>
```

Option H should specify information that is useful for understanding how to read DNA sequences, such as the meaning of each character representation of the nucleotides (see Reference).

Option S should display the name and identifiers for the sequence selected. It should also display the number of base pairs, the number of gap regions (specified by N's), the number of coded regions (specified by non-N characters, such as A,G,T,C, etc.), the number of gap region nucleotides, the number of coded region nucleotides, the number of A, G, T, C, etc. specified (see Reference for full list of nucleotide characters).

Option 1 should result in the user being asked for a number, n, which should represent the n^{th} gap region. Display the base pair number range, gap or coded region number, and the region's nucleotide symbols.

Option 2 should result in the user being asked for a number, n, which should represent the n^{th} coded region. Display the base pair number range, gap or coded region number, and the specified region's nucleotide symbols.

Option 3 should result in the user being asked for two numbers, n and m. Display the base pair number range and the nucleotide symbols specified by this range.

Option 4 should result in the user being asked for at least 10 nucleotide symbols. Display the base pair number range where these 10 nucleotides symbols occur in sequence. Display 20 symbols before, the nucleotide symbols specified, and 20 symbols after.

Option 5 should result in the user being asked for a file name that specifies at least 10 nucleotide symbols. The file should be in the FASTA format (`.fa` file name). Display the base pair number range where these 10 nucleotides symbols occur in sequence. Display 20 symbols before, the nucleotide symbols specified, and 20 symbols after.

Option R should return the user to the previous menu.

Option Q should quit the program.

For all menu options, the user should be asked to re-enter a value until a valid input is specified.

FASTA Data Files

Your code should only work for data specified in the FASTA format. For more information, visit:

https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=BlastHelp

One of the test codes in marking your work will use the human Chromosome 16, which has been uploaded onto blackboard. We have also supplied the sequence for the FOXF1 gene, which is located within Chromosome 16.

Useful C++ Standard Libraries:

```
#include <iostream>
#include <string>
#include <fstream>
#include <sstream>
#include <vector>
```

References

Character Specification

Code	Base Description
G	Guanine
A	Adenine
T	Thymine (Uracil in RNA)
C	Cytosine
R	Purine (A or G)
Y	Pyrimidine (C or T or U)
M	Amino (A or C)
K	Ketone (G or T)
S	Strong interaction (C or G)
W	Weak interaction (A or T)
H	Not-G (A or C or T) H follows G in the alphabet
B	Not-A (C or G or T) B follows A in the alphabet
V	Not-T (not-U) (A or C or G) V follows U in the alphabet
D	Not-C (A or G or T) D follows C in the alphabet
N	Any (A or C or G or T)

Genomes of Organisms

The following are links to complete genomes of organisms including homo sapiens.

https://www.ncbi.nlm.nih.gov/guide/howto/dwn-genome/

ftp://ftp.ncbi.nih.gov/genomes/

The assembled human genome was found in the following folder:

ftp://ftp.ncbi.nih.gov/genomes//H_sapiens/Assembled_chromosomes/seq

Chromosome 16 is compressed in the following file: “hs_alt_CHM1_1.1_chr16.fa.gz”.

Example Output

```
DNA Sequence Database Software

Specify the name of DNA sequence file names you would like to load. For multiple files, add a
',' between each file name.
>chr1.fa, chr16.fa

Loading chr1.fa...
Successful loading of chr1.fa
Loading chr16.fa...
Successful loading of chr16.fa

Select one of the following options:
(S) Summary statistics of the DNA database
(1) Analyse chr1.fa
(2) Analyse chr16.fa
(Q) Quit
>1

Select one of the following options
(H) Help
(S) Summary statistics of the DNA sequence
(1) Analyse gap region
(2) Analyse coded region
(3) Analyse base pair range
(4) Find DNA sequence by manual input
(5) Find DNA sequence by file input
(R) Return to the previous menu
(Q) Quit
>h

Code Base Description
G    Guanine
A    Adenine
T    Thymine (Uracil in RNA)
C    Cytosine
R    Purine (A or G)
Y    Pyrimidine (C or T or U)
M    Amino (A or C)
K    Ketone (G or T)
S    Strong interaction (C or G)
W    Weak interaction (A or T)
H    Not-G (A or C or T) H follows G in the alphabet
B    Not-A (C or G or T) B follows A in the alphabet
V    Not-T (not-U) (A or C or G) V follows U in the alphabet
D    Not-C (A or G or T) D follows C in the alphabet
N    Any (A or C or G or T)

Select one of the following options
(H) Help
(S) Summary statistics of the DNA sequence
(1) Analyse gap region
(2) Analyse coded region
(3) Analyse base pair range
(4) Find DNA sequence by manual input
(5) Find DNA sequence by file input
(R) Return to the previous menu
(Q) Quit
>s

Sequence identifiers:
Name:          Homo sapiens chromosome 1, alternate assembly CHM1_1.1, whole genome shotgun
sequence
GID:           528476670
REF:           NC_018912.2

Region characteristics:
# regions:    6305
# N regions:  3153
# C regions:  3152

Base pair characteristics:
```

```
# base pairs: 250522664
G:            47434131
A:            66127181
T:            66204528
C:            47449501
R:            0
Y:            0
M:            0
K:            0
S:            0
W:            0
H:            0
B:            0
V:            0
D:            0
N:            23307323
Unknown:      0

Select one of the following options
(H) Help
(S) Summary statistics of the DNA sequence
(1) Analyse gap region
(2) Analyse coded region
(3) Analyse base pair range
(4) Find DNA sequence by manual input
(5) Find DNA sequence by file input
(R) Return to the previous menu
(Q) Quit
>r

Select one of the following options:
(S) Summary statistics of the DNA database
(1) Analyse chr1.fa
(2) Analyse chr16.fa
(Q) Quit
>2

Select one of the following options
(H) Help
(S) Summary statistics of the DNA sequence
(1) Analyse gap region
(2) Analyse coded region
(3) Analyse base pair range
(4) Find DNA sequence by manual input
(5) Find DNA sequence by file input
(R) Return to the previous menu
(Q) Quit
>1

Enter gap region number:
>4

Selected sequence:
Base pair range: (214670,214940)
Gap region number: 4

Sequence:
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN

Select one of the following options
(H) Help
(S) Summary statistics of the DNA sequence
(1) Analyse gap region
(2) Analyse coded region
(3) Analyse base pair range
(4) Find DNA sequence by manual input
(5) Find DNA sequence by file input
(R) Return to the previous menu
(Q) Quit
>2
```

```
Enter coded region number:
>6

Selected sequence:
Base pair range: (238381,248227)
Coded region number: 6

Sequence:
TTTTTTTTTTTTTTGAGATAGGGCTTTCTTGTCACCCAGGCTGGAGTGCAGTGGCGGAATCAAGGTTTGCCTCAGTGTCCTGTGCTCAACAGAT
CCTCCTGCCTCAGCCTTCTTAGTAGTAGCTGGGACTACAGGCATGTGTCACCATGCCCAGCCCCATGGGTCTTTGTTTTCTGCTAGTAACTTCA
AAAGGGGACATGCAGTGTAACTCACCTGAGTCCCAAGTTAAGGCTATGGTAGCAGCTGAGCAGACAAGACAGCTTTTCCTAAATAAATGAATGC
AGAATGGGGACCCCTCCCTCCTCCCCATGGGAGATGGGCAGCTCAGACACTCAGAAGGTTGTGTGGGAGCGAGCAAACAAACACCCATGCACAC
CTGGGTCCCGTCAGTCAGATGCAGGGGACAGGAAGTGCCCTGCCATCTACTGGATGCCAGAAGACAAGACGTGACCCACGAGTAAGTCACGGTT
TCTGTGAGGTGCTGGTGGCACTGGCACAGGGTCACAGTGAAAAGCCTCTACGCAGAGGACAGCAGAAACCCCCCGCAGCCTCAGGAGGCAGCAT
CAGATTTATTTATTCCTACTCAACATGACCCGGGAACACAGGAGCAACTCTGTACACTTCTAGAAACTCACAGCTAGCTCCAAAACAATAGAAA
TTTTAAACTACAAAAGATGAGTTGTATTCAGCAAATATAAAGGGTAATTTTAGACTGTGTGAACGTTTATCAGACTATTTACAGCACCCGGGAG
ACGGGTTCAGATCTCGCCGGCCTCCTTCTCTTCTGACCTCCGTGAAGCCATCTTCCCGTTGGAGCTCTCAAGCCTCCAGTCCGGGGGCCCTCGC
TCGCTCCGCCCGCTCTCCCAGGACTCCTCTCTGGATGCCCGCTCTCTGGAGAACCTGGGAAATGGGAACCAGAGGCTCAGTGGAGGCTGCTGCC
CCCCTTCTGCTGGCCGCTCAGGCACTGACGATGGACCCGCAGACCAGGCCAGGCAAGGCCCTACATCCCCTATACCTGGAAACCCCCCGGAGGC
CAACCAGACTTGCCCACCACATATGGAGTGTTAGATAGCCATTAAAAAGTGACCTTTGCGAAGATTTTTAATAAGGAGAGCCTTCACTGTAACA
TAAACAGTGCAAAGGAGAATACAACACTATATGCACAGAACATGACACTACGTAAAAACACAATGGAAAAAAAAATATCAAAAACAAAAACCAA
AACAGAGGTAAGCAGGGCCTGGTCAGGACACAAGTCTCAGCCACTCTGAGCATGGACGCAACAGTGCACAGGCCCCAGGACGCAGGCAGGGGAC
AGTGCCTGTGCAGCCCATCACCTGGCGCCGTGCGAGCTCCTCCACAGCCCTCTCGCCTCTTCCCACACCTGAGTCCCGTTCACGACTCCCCAGC
CCTACTCCTGAGGGTGAGCCCAGCCCCATCCCCACCAGACACGGCCCTCAGCAGACCCTGGGAACACAGAGAAGGGCTCACTCCTGGTTGCAGC
TACCGAAGGAGTCAGAGTAGTTACTTGTATTTCGCACTTCGGTCCCGGGAAGCCCGACGATGTCCCCGGCTGTGGCTCCGGGAACGGCTGCGGT
GGCGCCGATGTCTATCTCGGGACCGGGACCGGGACAATTTCCGTCGCTCTCGGGAGGTAGATCTTGACCGCCTCCGACGCCGATCCCGGGAGCG
TGACCTGAACAATCAGAAGGCTCCAAGGCTGAGACTCTATAGGTGCCAACACTATGAGGGCCCCACCCCTCAGCAGATTCTGCTTGCTGCTTTA
CTGACATTTTCTCCTTTACTCCCACAAATTTAGGTTTAAAATATTTCCATCAGTTCATCTCCAATGGCTAGGTCCTTCCAATGAGGTCTCCTGA
AAACATGTGGCCTGGATACTCCTGTCACCCTGCAGGCCCCGACCTTCCCTCTCATTAACACTGGGAACCCTCCACGGCACCAAGCAACCTCTTG
CCCTGTGCCCTTCACAAGCCAGCCCTACCTGCCAGGTAACATTTGTAAAGGCATAATCATTAAATATTAAGAAAAATGCTTATGGAAAAAAAAA
TTTGAGAACATTTTAATGGACCTGGAAACCTACCTCCTGCGATCTCTGGTTCTTGATCCCGACCTAAAAGGGGGAAAAAAAGATGAACAAGGCC
AGGCATGGTGGCTCACACCTGTAATCCCAGCACTTCGGGGAGGCCTAGGCAGGCAATCACCTGAGGTCACGAGCTGGAGACCAGCCTGGACAAC
GTGGTGAAACCCCGTCTCTACTGAGAACACAAAACCTGGCCAGGCGTGGTGGCGGGTGCCTGTAATCCCAGCTACTCAGGAGACTGAGGCAGGA
AAATCACTTTAACCTAGGAGGCAGAGGTTGCAGTGAGCTGAGATTGTGCCACTGCACTCCAGCCTGGGCAACAAGAATAAAACTTTGTCTCAAA
AAGAAAAAAAAAAAAGTGAATGACTTCCCGAAAGCACCCCTATTCCTCACCCTCTTGGGGGAGGACTCCAAATTTAAGAAGTGAGGCACCTCCCC
CAAAGACCTGCTGTTTTCAAGCTGCCTGCATGTTCCCAGTCTGAATGGGAAGTGCCAATGCTCCCATCCCCACAGCCACTCTGTGGCACACTCA
GTGTCACCTGAGCCTCACAGACAGGCCCACATCCGAGCATCAGACGCAGCCCTGTAACATTACAAGGGCCGTGAAAAGTCTGTGCACACTGACT
GTGCTGCACTCAAGTCCTACAAATGATTCTCCAGTCTTTACAACAGACACCAGCACAGAATGTGCAGACAGCGAGTGGGGCACACCAAAAACAC
AGAAAAGGAAAAGCAAAGCCCGAAGTTCCAGGCACAGGCTTCCTGTCAGGTGAGAGTCATCAGCTCATTCTGCCACACAACCAAGAGGCTGAAA
AAGGCAAGCTATGAGGCCTCATTCTATGCCCGTAAGCCATTAAGGTCAAGTGGCCCCATGTACATCCGCAACACAAATGGCCTAAACATGACAG
TGCATGTAAAAATGTAAACTCCATGCAAAAAATGGATTGTGTGTTTTGCTTAATATTTTCTTCTATACTGTTTAAGTTTCTCTCCAATCAGAAT
CTCCTGCTGGTAAAACAACAAACGGCCGGGCGTGGTGGCTCACGCTTGTAATCCCAGCACTCTGGGAGGCTGAGGCGGGCAGATCATGAGGTCA
ACAGATCGAGACCAGCCTGGCCAACATGGTGAAACCCTGTCTCTATTAAAAATACAAAAATTAGCTGGGCGTGATGGCACAGGCCTGTAGTCCC
AGCTACTTGGGAGGCAGAGGCAGGAGACTCGCTTGAACCCAGGGGGTGGAGGCTGCAGTGAGCCGATATCGCACCACTGCACTCCAGCCTGGTG
ACAGAGCGAGACCCTGTCTCAAAAACAAACACAATAAACATTCGCCCTGAGTTATGCACTGAGGAATTATTTCAGGAAAAGTGGTTTGGGAACC
TTCCTGGACAATGAGCTGTCTCCAGGAGGATGTGGATACTTCTGGCTCCTACCTCTTTGAACTCCAAACAGACTCAACAGCCCCACCTCAAGAT
GGCCTCTCAAGCTGAGTGTGGCCAGAGGCATCTTCTACTTCCCACCTCCTCCCACTCCAATACACCACACCCAGCCTTCCAACCTCACCAAGCA
GCAACTCCGGGCCAGAAACCAGGAGACCAGCTTCATTTCCTCCTCTGCACGCTTCACACACTACCAGCAAATCCCATCAGTTCTATCCCAGAAC
TTTCCCCGACTCCTTGCACCCCTTACCTGGCTTAAACTGCCACCTCCTCCATCTGGATTTTCCCACCTAGCTCCCCAAGGTCCTGCTGCTTCTA
CTCCAGCCCAGCACGCTTGCCAAAGGGGAGCAGGCACCACCCCTCCTGCGTAAACAGGTGTCCAATGGCCCCATAGCACTCAGTAAGTAAACAG
CAGTGCCCACACACAGCCTCCAACAGCTTGGATCCCTTACCTGACTGCCATGCACACACAGTCCTGCGCTGGGACGTGTTTTGTTTTGTTTTGT
TGGACTGAATATGCTGTTTACTTTTTTTTTTTTTTTTTTTTTGGTTGAGATGGAGTCTCTGTTGCCCAGGCTAGAGTGCAGTGGCACGATCTCG
GCTCACTGCAACCTCTGCCTCCCGGGCTCAAGTGATTCTCCCACCTCAGCCTCCGAGTAGCTAGGATTACAGGCTCATGCCAGCACGCCCAGCT
AATTTCTGTATTTTTAGTAGAGACAAGGTTTCACCATGTTGGCCAGGCTGGTCTCGAACCCCTGACCTCAAGTGTTCTGCCCGCCTCAGCCTCC
CAAACTGACGGGATTACAGGCGTGAGCCGCCAGGCCTGGCCTTACTATTTTCTTTTATAACGGCCTATTGGGCAGGCCCTGCCTATTCCAAGCA
ACAGTGGAATGGACCGAGCAGGGAATGCCCGAGGCCAATGCAGGCCATCCAGTGCCAGCCCTCAGGCTCACCTCCTGCTCAGACGCTCCTCCCG
TTCCCTCTCCTCTCTCCTCCTCAAGCGATCCTGATTTCTCTTCTCCTGCTTTTCAGCGACAGTTTTCTAAATAAATGAAACAAAAAATGAGGAA
GAGCAAGTTACACAAACCAGAGTGTGAATGCTACTTAACAAATCACCTTTATATGAGCTCAGTATTGGTAATTTAAAAATTGAAGGGACACTTT
TAGGAAATGGGAATACTTTTTTTTTTCTCTTTGAAACGGTGTCTCACTGTCACCAGGCTGGAGTGCAGTGGCGTGATCTCGGCTCACTGCAACT
TCCACCTCGAGTTCAACCTATTCTTCTGCCTCGGCCTCCCAAGTAGCTGGGATTACAGGCACATGCCACCACACCCAACTAATTTTTGTATTTT
TAGTAGAGATGGGGTTTCACCATGTTGGCCAGGATGGTCTTGATCTCTTGACCTTGTGACCCGCCTGCCTCAGCCTCCCAAAGTGCTGGGATTA
CAGGCATGAGCCACCGCGCCCGGCCAACACATTCATTTTTAAAACTTCTTAATGAAGAAGATAACACATGAAAAAATGCTTTATCATTACATTT
TTTTTTCCTGAGACAGAGTCCGCCCTGTTGCCCAGTGCAGTGGCGCGATCTTTGCTCACTGAAGCCTCTGCCTCCTGGGTTCAAGCGATTCTTC
TGCCTCAGCCTCCTGAGTAGCTGGGATTACAAGCGTGCCAGCACGGCCGGCTAATTTTTGTATGTTTAGTAGAGACGGGGTTTCACCATGTTGG
TCTGGCTTGTCGCAAACTCCTAACCTTGTGATATGCCTGCCTCGGCCTCCCAAAGTGCAGCAATTACAGGCGTGAGGCACTGCACCCGGCCTCG
TTTACATTTACTTTTTTTTTTTTTTTTGAGACGGAGTCTCGCTCTGTCACCCAGGCTGGAGTGCAGTGGCGCAATCTCGGCTCGCCGCAAGCTC
CGCCTCCCGGGTTCACGCCATTCTCCTGTCCTGGCCTCCGGAGTAGCTGGGATTACAGGCGCCCGGCACTGCGCCCGGCTAATTTTTTGTATTT
TTAGTAGAGACAGAGTTTCACCATGTTAGCCAGGATGGTCTCGATCTCCTGACCTCACGATCCGCCCGCCTCGGCCTCCCACAGTGCTGGGATT
ATAGGCATGAGCCACCGCGCTTGGCCACTTTTTTTTTTTTGAAACAGGGTCTTGCTCTGTCCACCCAGGCTGGATGGAGTGCAGTGGCGCAATC
ACAGCTCACCGGAGCCTTGACCTCCCAGGCTCAAGCAAGCCTCTTTAGCTCCGCCTCCTGAGTAGCTGGGACTATAGGCGTGTGCCACCATGGC
AGACTAATCTTTAAACATTTTTTATGGAGATGGAATCTTGCTATGTTGCCCAGGCTGAGCTTGAACTCCTCAACTCAAGCAATCCTCCTGCCTT
GGCCTCCCAAAGTGCTGGGGTTACAGGCCTGAACCACCCAGCCTATCATTAGATTTTAACAATGTGGTAACAAACTGTCTATTCTAAGTGGTGC
TATACTTCTAACTCACAGGGAATTGATCACATTGTTTCAAACCAACCCAGATAGGTAGGAGACTGTACAAACTTTATGGTAAAAAAGGACGGAT
```

```
GTATCTCATCAAACACATCTTGCTACTGGGAGCAGAATCCCTACAGCCAAAAAGCAGCCCAACCTCAGGGTGCTGGGTGCTACGCTGGCTTTAG
AACAAACTGAGTTGGAGGACAGCAGGCCAGGGGAGGTGGACCACAAATCCTCTGCAACTGGCTTTCAACACAGGATGATGTGAATCATCAGGCT
TGATGAAGAAAGCAATATAAGGGGATGAGAAACCTCGCCAATGAAAATACCGTCAACACAAGCTGAGTGGAAAAGCACCGGGCACCGTTCTGTA
TACCCACTACTCCCAATATCAACTCTAAGACATTTAATGCCCCTCATCTGTGTCAAGTGAAGCCCTGAGTGGGAGCAAGTCAGAGCCAGTAAGA
CAAAGAATCACGAGTCTTCTCCATCCAGAGAGCCACGATGGCATAAGGATAATCTTCCATACCCAGTGTGGATGAGAGCACGCTCCCCACAAGG
GCACGGAAAGTATGCAGAGCTGGGTGACTGGTATCTCCAGCTCTAAACTTTCACCCATGCTCACCCTGTGTCTGAAGCAGGGCACACATGGATG
GACACCTTTAAGGTGATGTTTAAGGATGGCATTTAAGCCAGGTACGGTGGCTCATGCCATTTCAGCACTCTGGGAGGCTGAACTGGGAGGATCA
TGTGAGGACAGGAGTTTGAGACCAGCCTGGGAAACATAGTAAGACCTCCTACCTAAAAAAAAAAAAATGAGCCAGATATGGTGGTGCATGCCTGT
AGTCTCTGCTACTCAGGAGACTCAGGATGCAGGGCCACTTGAGTCCAGGAGTTCCAGGTTACAGTGACAATAAGCTATCAGCTACCATCACATC
AGTACACTCCACCCTGGGTGCCAGAGGGAGACCCTGTCTCTGAAAAAAAAAACGAAAAAAATCCTCTGTTGTCTAGGTTGGAGTGTAGTGTCGT
GATCACAGTTCACTGCAGCCTTGAGCTCCAAGGTCCAAACAATCCTCTTACCTCAGCCTCCTGAGTAGCTGGACCTACAGGCACACACCACCAT
GCTTGGCTAATTTTTAAATTTTTGGTAGAGACAGGGTCTCCTTATGTTGCCCACGCTGGTCTCAAACTCCTGGATTCAATCGATTCTCCTGCCT
TGGCCTCCTAAAGACTGGAATTACAGGCGTGAGCCACCACACCCAGCCTCAAGGATGGCAGTTTTTAAAAAAATTAAGGATTAGCCAGGTGCAG
TGGCTCATGTGTAATCCCAGCACTCTGGGAGGCCTAGGAAGGTGGATACCTCTGGGACTACTGTGAGTGAGGCCTCCTCTAACTAGGGCACAGA
ACTCCCACTCTCCCTATCATTAAGAGGTGCACTTCCAAATACATTGTGCACTGTTACTTTAATAATTATTTATCAAAGTCTGTAATACGAGTTG
TGTTGTTTTGGTCAACTCTATGCTGCAATGATAAAACAAGCTTTCTAACCAAGTTAATGCTTCTTAAGGTAGAAGGGAAAAACATTTCACAATA
TTCACATTTCTAATGCAGAGATACACATTAAGGAAATAAAAGATTGCAGATATTAAAAAAAAAATACTACATTAGGGCCAAGCACTGTGGCCCA
CGCCTGTAATCCCAGCACTTTGGGAGGCCGAGACAGGTGGATCACCTGAGCTCAGGAGTTCGAGATCAGCCTGAGCAACACGGTGAAACCCCAT
CTCTACTAAAAACACAAAAATGAGCCAGGCATGGTGGTGCACGCCTATAGTACCAGCTACTTGGAGGGCTGAGGCAGGAGTATCTCTTGAAGCT
GGGAGGTGGAGGTTACAGTGAGCTGAGTTCGTTACACTGCACTCCAGCCTGGGTGACAGAGCAAGACCCTGCCTCAAACAAACAAACAAACAAA
CAAAAAAAAACCAACAACTACAGTAGGACAAAATTCTGAGCTGTGAAATGAAAATGTAATTTAAAGGAGAAATGGGAAAGTAACATTTCCAGTG
GCTAACAATTTTTTTTTCTTTAAGACAAAGTCTCACTCTGTTGCCCAGGCTAGAGTGCAGTGGCTCGATTTCGGCTCACTGCAACCTCTGCCTC
CCGGGTTCAAGCGATTCTCCTGCCTCAGCCTGCCAAGTAGCTGAGAATACAAATGTGTACCACCATGCCCAGCTAATTTTTGTGTTTTCAGTAG
AGACAGGGTTTTGCTATGTTGGCCAGGCTGGTCTCGAACTCCTGACTTAAGTGATCCGCCCTCCTCAGCCTCCCAAAGTGCTGGGATTACAGGT
GTGAGCCACCATGCCTGGCCAAAAGCAAATTTTTAATAGAGTGGTAAATGCATACTAAATTGTGATTTATAATTCTATTGGATACAGATACTTT
TACATCTTTTTTTTTTTTTTTTTTTTGAGACAGAGTCTCGCTCTGTCGCCCAGGCTGGAGTGCAGTGGTGCAATTTCGGCTCAGTGCAAGCTCCG
CCTCCCGGGTTGACACCATTCTCCTGCCTCGGCCTCCCGAGTAGCTGGGACTACAGGTGCCCACCACCATGCTCGGCTAATTTTTGTATTTTTA
GTAGAGACGGGGTTTCACCTTGTTAGCCAGGATGGTCTCAATCTCCTGACCCAGTGATCTGCCCGCCTCGGCCTCCCAAAATGCTGGGATTACA
GGAGGCTGAGCCTCTGCACCCAGCCTTTTTTTTTTTTTTTTTTTTTTTTTGAGACGGAGTCTTGCTCTGTCTCGTAGGCTGGAGTGCAGTGGCAC
AATTTCAGCTCACTGCTACCTCTGCCTCCCAGGTCAAGCAATTATCCTGCCTCAGCCTCCCGAGTAGCCAGGATTACAGGCGCCCGCTGCCATG
CCTGGCTAATTTTTGTATTTTTAGTAAAGACGGGGTTTCACTATGTTGGCCAGGCTGGTCTTCAACTCCTGACCTCGTGATCCACCTGCCTCGG
CCTCCCGAAGTTTGGGATTACAGGCGTGAGCCACCACACCCAGCCTCACTTTTACATCTTTCATACCTATAAACTTATAAAGAAAAATTCCAAC
TTTGACAGAGTACATTGCTTTTCAAAATTCTCTCGGGAACACGGAAGTAAAATGTTTCAAGACCAGTGGCCGAGAATAACCTCACTGCTGCACC
AGCTCCCATCACAGTCCCAGGTTCACACCATCACCTATGGGACCCTCCTGAGGTGCACAGCCCTGCCATATGCGGCCTGAGTGAGCAGCTCTTT
CCTGGGGCACAAAAGGCCCTCAGCCGTGGTCACTGCAGCATTCTGCAGTCAGCGACACGTCTGTCCCTCCACTACATTTCAGGTATACAGGGGG
AGGCTGGCTAATTTCCCACCATCCTATCCTGAGGTCCTAGTATGCGGTGAGGGGGGCTCCATCAGTGTGAAGGGAGACAATTCAAAGTAGTGGA
AATATGTCCAACCTTCCTTCTTAGCATTCCGTACTTAAAGTTTTTAAATGTCGGCCGGGTGCGGTGGCTCACGCCTGTGGTCCCAGCACTTTGG
GAGGCTGAGGCGGGCGGATCGCGAGGTCAGGAGATCGAGACCATCCTGGCAAACACGATGAAGCCCCGTCTCTACTAAAAATACAACAACAACA
ACAAAAATTAGCCGGGCGTGGTGGCAGGCACCTGTAGTCCCAGCTACTGTGGAGGCTGAGGCAGGAAAATGGTGTGAACCTGGGAGGCGGAGCT
TGCAGTAAGCCGAGATTGGGCCACTGCACTCCAGCCTGGGCGACAGAGCGAGACTCCATCTCAAAAAAAAA

Select one of the following options
(H) Help
(S) Summary statistics of the DNA sequence
(1) Analyse gap region
(2) Analyse coded region
(3) Analyse base pair range
(4) Find DNA sequence by manual input
(5) Find DNA sequence by file input
(R) Return to the previous menu
(Q) Quit
>3

Enter a comma ',' separated base pair range:
>2000, 2200

Selected sequence:
Base pair range: (2000,2200)

Sequence:
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNN

Select one of the following options
(H) Help
(S) Summary statistics of the DNA sequence
(1) Analyse gap region
(2) Analyse coded region
(3) Analyse base pair range
(4) Find DNA sequence by manual input
(5) Find DNA sequence by file input
(R) Return to the previous menu
```

```
(Q) Quit
>4

Specify the DNA sequence nucleotides you would like to find:
>GCTAAATGTCAT

1. Base pair range: (6444760,6444771)
AAATAAAATGTAAAGTTAAA
GCTAAATGTCAT
TGTATTCTATGCTTGCTGTA

2. Base pair range: (6718584,6718595)
CAGGAAAAACCTTGGCCACT
GCTAAATGTCAT
CCTGTGATATATCTCTGCAT

3. Base pair range: (10161028,10161039)
AGCTGCCCCTCCCAGTGTTT
GCTAAATGTCAT
GGCAGAGGGTTAACCTCCCA

4. Base pair range: (27830827,27830838)
AATGAAACATTGTTATTGTG
GCTAAATGTCAT
TGCTAGATAAAAGCTGTGAG

5. Base pair range: (48836846,48836857)
ACCATGTAACACAATCTGCT
GCTAAATGTCAT
ATCAGTCTCTTGGTAAAAGC

6. Base pair range: (62734171,62734182)
AGCAAATACAGCTATAAAAA
GCTAAATGTCAT
GGAGTTGTAGAAAAAAGTTA

7. Base pair range: (73701719,73701730)
AAAGTAGATTTCTAAAAATA
GCTAAATGTCAT
TTGGGATCAAACTTCTAAAA

8. Base pair range: (78849904,78849915)
CTTAGAAAATTTACTATGGG
GCTAAATGTCAT
CTAAGAATCAACACTGGAAA

Select one of the following options
(H) Help
(S) Summary statistics of the DNA sequence
(1) Analyse gap region
(2) Analyse coded region
(3) Analyse base pair range
(4) Find DNA sequence by manual input
(5) Find DNA sequence by file input
(R) Return to the previous menu
(Q) Quit
>5

Specify the DNA sequence file you would like to find:
>gnFOXF1.fa

Loading gnFOXF1.fa...
Successful loading of gnFOXF1.fa

1 matching sequence(s) found.

1. Base pair range: (87955849,87959786)
CCTCCCGGCCCGTCCGCGGC
GCAGAGCAGCGGCGGCAGCGGCGGCGGCGGCAGCAGCCACCCGATGTCTTCGGCGCCCGAGAAGCAGCAGCCACCGCACGGCGGCGGCGGCGGC
GGCGGCGGGGGAGGCGGCGCGGCCATGGACCCCGCGTCGTCCGGCCCGTCCAAGGCCAAGAAGACCAACGCCGGCATCCGGCGCCCGGAGAAGC
CGCCCTATTCCTACATCGCGCTCATCGTCATGGCCATCCAGAGTTCACCCACCAAGCGCCTGACGCTGAGCGAGATCTACCAGTTCCTGCAGAG
CCGCTTCCCCTTCTTCCGGGGCTCCTACCAGGGCTGGAAGAACTCCGTGCGCCACAACCTCTCGCTCAACGAGTGCTTCATCAAGCTACCCAAG
GGCCTTGGGCGGCCCGGCAAGGGCCACTACTGGACCATCGACCCGGCCAGCGAGTTCATGTTCGAGGAGGGCTCCTTTCGGCGGCGGCCGCGCG
```

```
GCTTCCGAAGGAAATGCCAGGCGCTCAAGCCCATGTACAGCATGATGAACGGGCTCGGCTTCAACCACCTCCCGGACACCTACGGCTTCCAGGG
CTCGGCCGGCGGCCTCTCGTGCCCGCCCAACAGCCTGGCGCTGGAGGGCGGCCTGGGCATGATGAACGGCCACTTGCCGGGCAACGTGGACGGC
ATGGCCCTGCCCAGCCACTCGGTGCCCCACCTGCCTTCCAACGGCGGCCACTCGTACATGGGCGGCTGCGGCGGCGCGGCGGCCGGCGAGTACC
CGCACCACGACAGCTCGGTGCCCGCCTCCCCGCTGCTGCCCACCGGCGCCGGTGGGGTCATGGAGCCGCACGCCGTCTACTCGGGCTCGGCGGC
GGCCTGGCCGCCCTCGGCGTCCGCGGCGCTCAACAGCGGCGCCTCTTATATCAAGCAGCAGCCCCTGTCCCCCTGTAACCCCGCGGCCAACCCC
CTGTCCGGCAGCCTCTCCACGCACTCCCTGGAGCAGCCGTATCTGCACCAGAACAGCCACAACGCCCCAGCCGAGCTGCAAGGTGAGTGGGGAG
GCCGAGGGCGCCCTGGTCCCCGGGAAGTCGAGTCTGAGTGGCAGCGGGACCCAGCTGGGGCGAGCCCCTCCACTTCTGTGGTCGGAACCCCAAG
GCTGAGGGGAGGCACCAGCTCCCCAAGGTGTCTCTTGGCCCCACCTCTCCCCCTTCGAAGAGTGACTACCGCTCTTGACCCTAGTTTGGGCCAA
TCTGTTTCTCTTTCTGTTCGAACTCCAGCTGCCAGCTGCTCCAGGCCTGACCAGGTAGGCTGGTCCTGAGCCAGATGTCCCAGACCCCAGTGCC
CTAAGTCGTTTTGTGTCCCTTAAGTCCCCTCACAGCTTGGAAAGATCCTGGGATGGACCCAAGGCTCCCAGCGCTGGCCAGATGGCTGTCCCTC
CCTGGTGGTGGCCCTCAGGCTGCCTGGAGGGCTGCCTCTGCCTGGGGCTGGGCGGAACGGAAGGTGTTAGGCCCAAAGCCGCTGGGCCAGCGCT
CAGAGGCTCAGCCCCGGCCTTTCAGGGGACCAGACCTGGCAGACCCACAGCTGAGTCCAGGGGCATTTTGTGGTGACTGAGGCAGGAGGGTGCC
GTGGCCAGGCCCCGGAGGGTCAGAGCTGCAGACCGGGCCATGGGGGCTCCTTGTTTTTTAGGATCCAAGTTTTCAGGTCCCCACACCCCCAACA
CCCCTGATACCCCAATACCCCGAACATGCAGAAGCATCTGCCAAGGGAGATGCAGCAGCCTCGCCTCCTGATTCTGCCCCAGGCCCCCCGCAGC
CTGCAGGCCCAGCTCTGGAAAAAGCTGGAGTGAGCTTGCCAGGGCCAGTCCAGGGTCAGCTGTCAGCAGCCCCAGGAGGCAGGAGCCAGGCCTG
GAGGCCTTAGGCAGCCTCTCTGCGGGGGACTCCCCGGCTCCGGGGCCTGTTCTCTGGAGAAGCTTCTTCTCACCCCTGCTGCTGGCCCGGCGCT
GGCCCTGCACTGGCCTCTCCTTCAACCCCGCCGGCAGTTGTAGCCTCTGAGGTTGGGGGTGGTGGTAGCTGCCCTGGCGGTGACGGCCATGGGC
GTTAGGGGCCCTCCACTTGGTCTCCCTCACCACTTCCCCTATGGCTCTGGAAGCCCTGGGCCCCTGCACGGTGTCCGAGGCTGGACAGGCACGC
AGCAGGCAGGCAGGCAGGCAGGCAGGCCGTGGCAAGTGTTCTGGAGTCTTTTCCTTGGGTGCGCAGCTGGAAGGCCGGGACTCGGGGAGG
AGAGCGGGGAGGCGGCCGTGCTCTGGAGCCAGGCTGACAGGCCCTGTGCGCCCCACGGGAGCACTGCTCCTCTGCCTGAACTCTGAGCCACCGT
GGCTAACTCTTCTGCTCCCCCAACCCCTCCTGTCGCCTCGCCTTGCAGGCATCCCGCGGTATCACTCGCAGTCGCCCAGCATGTGTGACCGAAA
GGAGTTTGTCTTCTCTTTCAACGCCATGGCGTCCTCTTCCATGCACTCGGCCGGCGGGGGCTCCTACTACCACCAGCAGGTCACCTACCAAGAC
ATCAAGCCTTGCGTGATGTGAGGCTGCCGCCGCAGGCCCTCCTGGTGCAGGCAGGCGGGTCACAGGGACCCTGGACCGGCACAAGAAACTGCTT
TCTTCTCGAGGTATAACCGTCGGCAGAAGAAAAGGGTTCCACCTCTCCCCAACCGGAGTTTTTGGCAAGGAGTCCCCAATGCAAAGACACAGCG
CTGCGGTTGGCACCTCCTTCCTCACTCCCTCAAAATTGTTAAGAAATGTTAGTGGTGGGTCTGATCTGACTGCAGCCATCGGTAAATAAAAGTT
TTTGATCCTGTTGAACCCGCCTGAGACGGTGCTGTGCAGGGGAAAGCCCCCGCACCCACACAGGAATTCTGCTGAGGTCCCCCCTCCTTCCGGC
CAATGGCAGAAGTGGGGGAAAATTTTTAGAAGAAAAGCAAACATGTGAGACCAATCATTATCAAATACTTTTATTTTTTGGTTGAGTATTTATC
TTTTTATTTTTTATTTTTTTTTTTTGAAAGAATGTCTTGGAATGCGCAAGTCTCCCTTTAGAGCCGTCTTTTGCAGGGAGCGGGAAGTGACAAGA
GCTCAGATCTCCCTCCCGATCTCCCTCCCCACCTCCGAAGTCTCCTCCGTGGACCACAGGTGGATCTTTGTGCGAACAACTTGCATTTCGGAAG
CCACTGTCCGTCTTTAAACAGAAAGTCGAAGGAGCCACGAAGCAAGCGGCCGTCCGGGCGTCCGCCTCCGTCCCCTTCCATGTTCCTCCTCTTC
CTTCGCTTCAGCCTCTTCTGTTATGTTTTGTCTTGAATTTTATTTAGACTTTTTCAGTGGGTATTTTTCTGTCTTCCAACCTCTACTGTAAACT
TTCTGGTCCGAGAACGAGCCGAACACAGCGCGACGCAGGGACTAGGACGGCCCGGTGACCGCGCGGATTCAGGATTGCGGGGACGCAGAAAGGT
TAAGGCACTTTTAAAAACTATAGCAAGGCTCCTGTTTATTTATTCTACTTTCTTTCCCTAATAATCAAAACACCGCGTAGGCTCCTCCGTTTAT
CAGTATTAATGGTGTAACTTTGTTGGCAATATTTGCCGTGTAGAATTTTTTTTAGATATCCATTGTAAATTTGAAACAAAGACCGATCTGTGTA
AAAACAAATTTCCATATGTTTTATATAAATATATATATAATATGAAGGACTACCCTCCTTTTTTTTTTTGTATTTTGGCTGCTAGAGTGCAGCA
TTTGTGACACGTATTTGAAATTTGAAATTTCCTTCTGCACTGTATAAAAGGACCATTTGAGGATGTTTTGCCTTTTGTGTATTTTTTCCTAAAA
AAAGAACAAAAATAAAAATGTATAACATTTGTACATGGCCTTTAAAATTGTATCAACTAGAAATAAAATTGCATGAGTATTTTA
TTGGGACTGAGATTGTAGAA
```

```
Select one of the following options
(H) Help
(S) Summary statistics of the DNA sequence
(1) Analyse gap region
(2) Analyse coded region
(3) Analyse base pair range
(4) Find DNA sequence by manual input
(5) Find DNA sequence by file input
(R) Return to the previous menu
(Q) Quit
>r

Select one of the following options:
(S) Summary statistics of the DNA database
(1) Analyse chr1.fa
(2) Analyse chr16.fa
(Q) Quit
>s

The DNA Sequence Database holds 2 sequences.

Sequence 1:
Name:          Homo sapiens chromosome 1, alternate assembly CHM1_1.1, whole genome shotgun
sequence
GID:           528476670
REF:           NC_018912.2
# base pairs: 250522664

Sequence 2:
Name:          Homo sapiens chromosome 16, alternate assembly CHM1_1.1, whole genome shotgun
sequence
GID:           528476567
REF:           NC_018927.2
# base pairs: 91765909
```

```
Select one of the following options:
(S) Summary statistics of the DNA database
(1) Analyse chr1.fa
(2) Analyse chr16.fa
(Q) Quit
>q

Program ended.
```