Index - 15743

Motif Finder

Introduction

This Motif Finder which is also a Sequence Analyzer is a versatile tool designed for bioinformatics analysis of DNA, RNA, and protein sequences. With this tool, users can perform a variety of sequence analysis tasks, including identifying sequence type(DNA, RNA or Protein), specific sequence elements such as TATA and CCAAT boxes, calculating AT and GC content, searching for motif patterns using regular expressions in a given sequence, and predicting the presence of promoters in DNA sequences based on TATA and CCAAT motifs and GC content.

Features and functionalities

1. **Sequence Analysis Options**

The tool offers a menu-driven interface that allows users to select from different sequence analysis options, making it easy to perform specific tasks according to the user's requirements.

1.Sequence Element(TATA/CCAAT Box, AT Content, GC Content)

TATA box : This option allows users to search for the TATA box motif within DNA sequences. These are conserved nucleotide areas seen at close to 25-30 base pairs upstream to the site of transcription initiation. Consensus sequence – TATAAA

CCAAT Box: This option enables users to search for the CCAAT box motif within DNA sequences. These are conserved nucleotide areas seen at close to 75-80 base pairs upstream to the site of transcription initiation. Consensus sequence – GGCCAAT

AT Content - Users can calculate the AT content of DNA sequences. AT content refers to the proportion of adenine (A) and thymine (T) bases in the DNA sequence relative to the total number of bases.

GC Content - This option allows users to calculate the GC content of DNA sequences. GC content represents the proportion of guanine (G) and cytosine (C) bases in the DNA sequence relative to the total number of bases.

2.Motif pattern

The tool employs regular expressions to search for motif patterns within sequences, enabling users to identify specific sequence patterns of interest in a given sequence(DNA, RNA or Protein)

3.Promoter prediction

For DNA sequences, the tool can predict the presence of promoters by analyzing the arrangement of TATA and CCAAT boxes and the GC content within the sequence.

4.Sequence type

Identify the type of sequence. It checks the characters present in the sequence against predefined alphabets for DNA, RNA, and protein. If the sequence contains characters consistent with one of these alphabets, it is classified accordingly.

2. **Support for Multiple Sequence Types**

It supports analysis of DNA, RNA, and protein sequences, adapting its functionality based on the type of input sequence provided by the user.

3. **Flexible Input**

Users can input sequences interactively or read sequences from a FASTA file, providing flexibility in how sequences are processed.

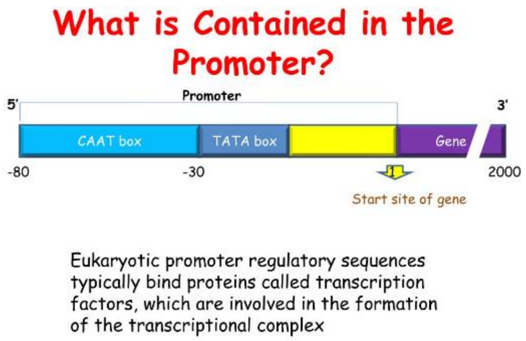
4. **User-Friendly Output**

Results are presented in a user-friendly format with colored output, enhancing readability and facilitating interpretation of the analysis results.

5. **Modular and Extensible**

The code is well-organized into classes, making it modular and extensible. Additional functionalities can be easily added by extending existing classes or creating new ones.

Promoter prediction Criteria



What is Contained in the Promoter?

Eukaryotic promoter regulatory sequences typically bind proteins called transcription factors, which are involved in the formation of the transcriptional complex

CAAT-Box. Elaine Chiu Eden Maloney Nancy Phang. Transcription

https://www.slideserve.com/meghana/caat-box

Generally, TATA box motif seen at close to 25-30 base pairs upstream to the site of transcription initiation and CCAAT Box motif seen at 75-80 base pairs upstream to the site of transcription initiation. CCAAT Box should be upstream from the TATA Box and gap between two should be 50 (between 40 – 60).

In eukaryotic promoter regions GC Content should be higher. (Generally in vertibrates, 40-45%)

The GC threshold is set as 30% for eukaryotes. (This also may differ with species as this value is highly species specific and hard to give a specific threshold.) (References)

Limitations

Limited Generalization: The GC threshold and motif-based approach used in the tool may have limited generalization capabilities. They rely on GC threshold, predefined sequence motifs and distance constraints between motifs, which may not capture the full complexity and variability of promoter sequences across different organisms or genes.

The method heavily depends on the presence of specific sequence patterns, such as the TATA box and CCAAT box motifs, to predict promoter regions. However, not all promoters contain these canonical motifs, and there can be significant variability in promoter sequences, making it challenging to accurately predict promoters solely based on motif presence.

The method used in the motif finder tool offers simplicity and computational efficiency, it may suffer from limitations in accuracy, generalization, and adaptability compared to ML-based approaches for promoter prediction.

ML approaches offer a data-driven and flexible framework for promoter prediction.

Future Improvements

Enhanced Motif Search Algorithms: Implement more sophisticated algorithms for motif searching, such as Position Weight Matrices (PWMs) or Hidden Markov Models (HMMs)

Explore the use of machine learning algorithms for promoter prediction.

Extend the tool to support analysis of additional sequence types.

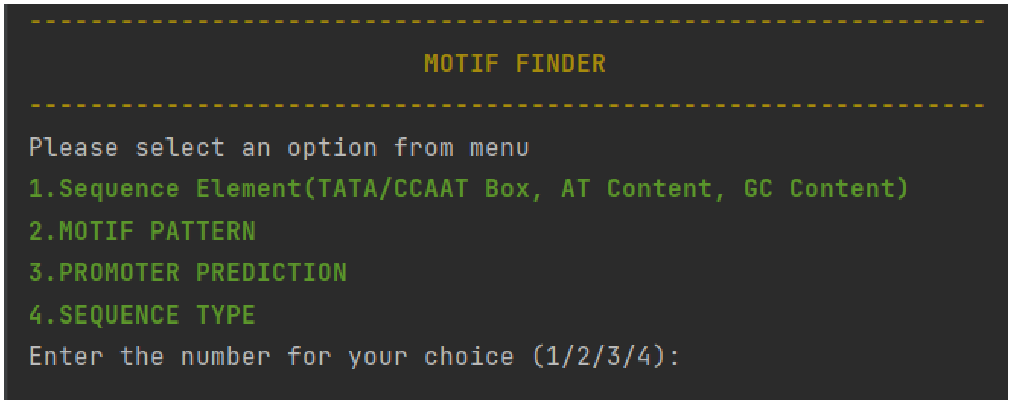
Incorporate integration with external databases and resources, such as NCBI or UniProt, to retrieve additional information about identified motifs, including functional annotations.

Integrate interactive visualization capabilities to display sequence motifs and their positions within the input sequences. Interactive plots or graphical representations can aid in the interpretation of results.

Index - 15743

How Application Works

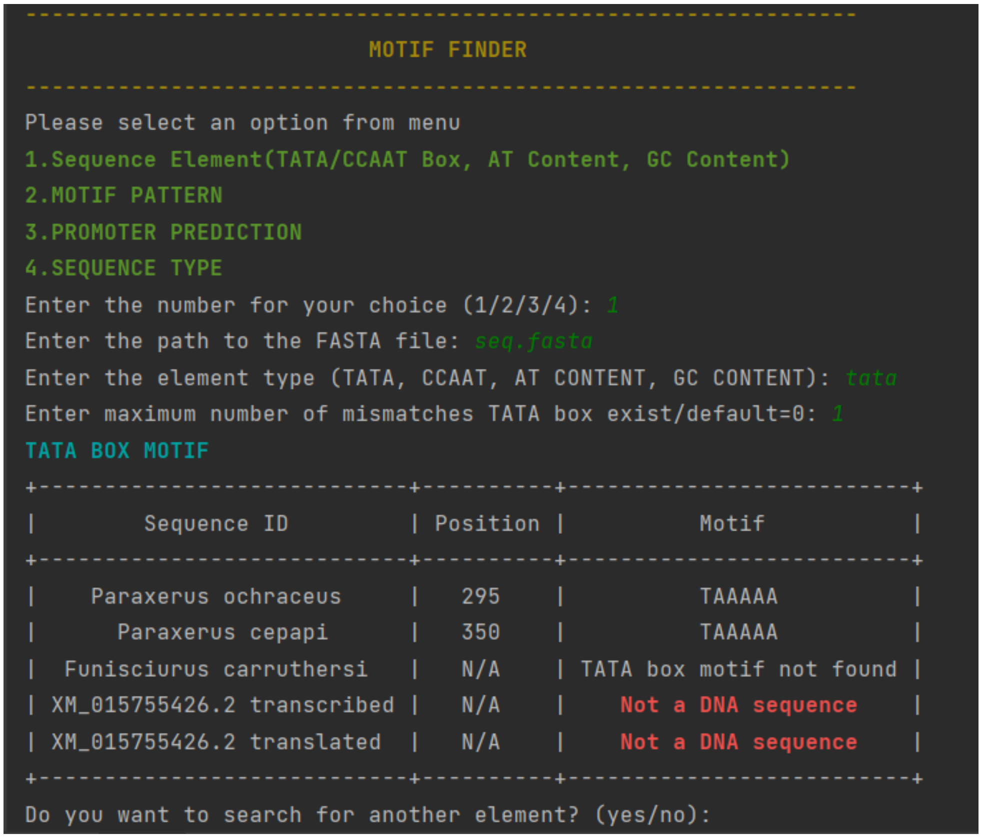
Main Menu



```
---------------------------------------------------------------
                         MOTIF FINDER
---------------------------------------------------------------
Please select an option from menu
1.Sequence Element(TATA/CCAAT Box, AT Content, GC Content)
2.MOTIF PATTERN
3.PROMOTER PREDICTION
4.SEQUENCE TYPE
Enter the number for your choice (1/2/3/4):
```

Sequence Element Search

TATA Box

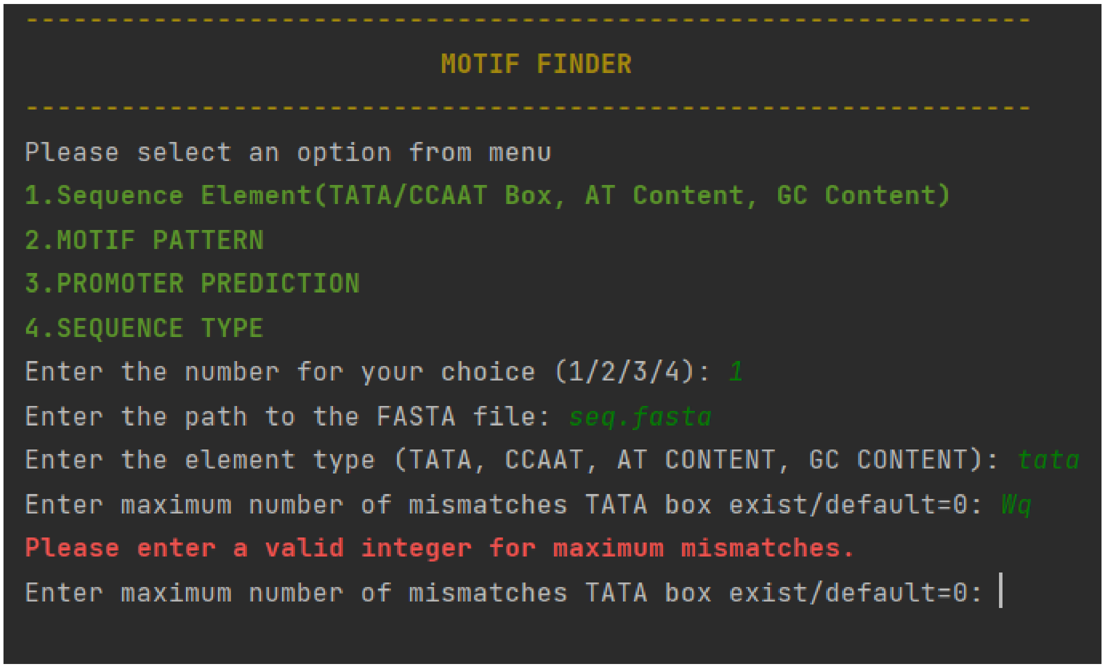


```
---------------------------------------------------------------
                         MOTIF FINDER
---------------------------------------------------------------
Please select an option from menu
1.Sequence Element(TATA/CCAAT Box, AT Content, GC Content)
2.MOTIF PATTERN
3.PROMOTER PREDICTION
4.SEQUENCE TYPE
Enter the number for your choice (1/2/3/4): 1
Enter the path to the FASTA file: seq.fasta
Enter the element type (TATA, CCAAT, AT CONTENT, GC CONTENT): tata
Enter maximum number of mismatches TATA box exist/default=0: 1
TATA BOX MOTIF
+----------------------------+----------+-------------------------+
|         Sequence ID        | Position |          Motif          |
+----------------------------+----------+-------------------------+
|    Paraxerus ochraceus     |   295    |         TAAAAA          |
|      Paraxerus cepapi      |   350    |         TAAAAA          |
|  Funisciurus carruthersi   |   N/A    | TATA box motif not found |
| XM_015755426.2 transcribed |   N/A    |    Not a DNA sequence    |
| XM_015755426.2 translated  |   N/A    |    Not a DNA sequence    |
+----------------------------+----------+-------------------------+
Do you want to search for another element? (yes/no):
```

Input – Fasta file containing 3 DNA sequences, 1 RNA sequence and 1 Protein sequence.

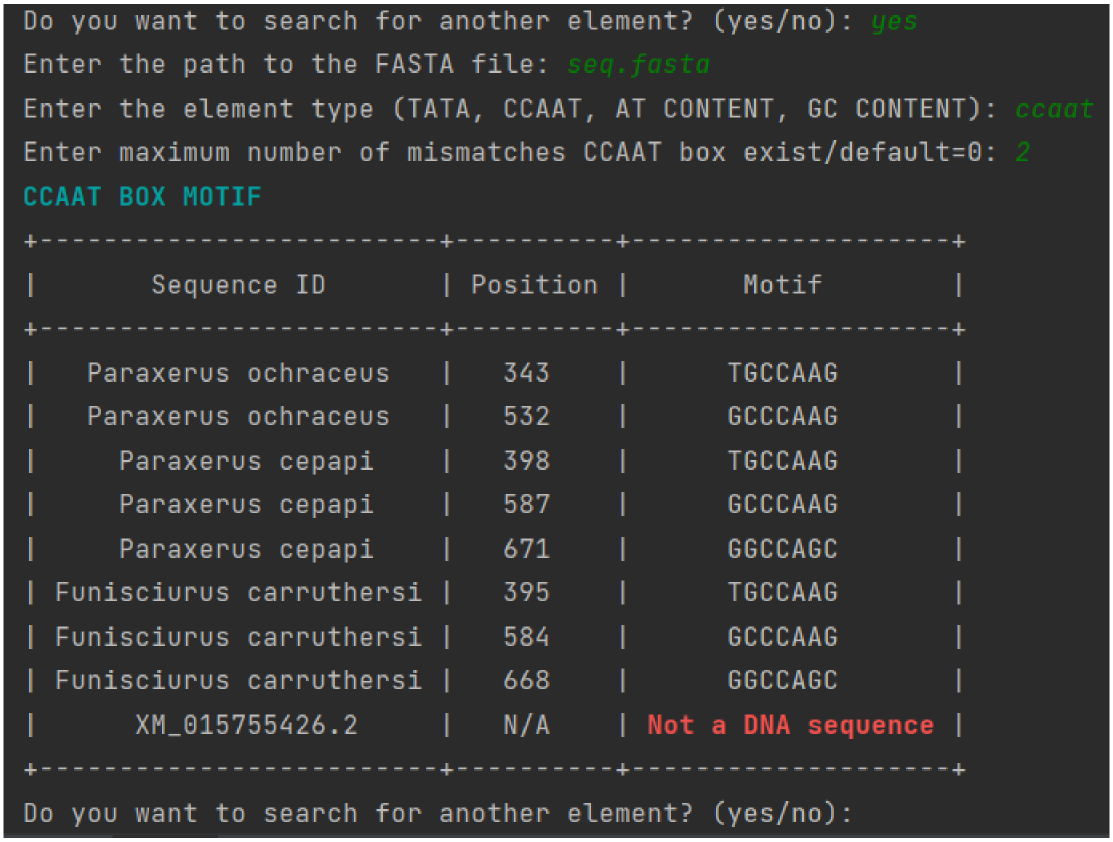
Index - 15743

Incorrect input to max mismatches (i.e. non integer input)



```
---------------------------------------------------------------
                          MOTIF FINDER
---------------------------------------------------------------
Please select an option from menu
1.Sequence Element(TATA/CCAAT Box, AT Content, GC Content)
2.MOTIF PATTERN
3.PROMOTER PREDICTION
4.SEQUENCE TYPE
Enter the number for your choice (1/2/3/4): 1
Enter the path to the FASTA file: seq.fasta
Enter the element type (TATA, CCAAT, AT CONTENT, GC CONTENT): tata
Enter maximum number of mismatches TATA box exist/default=0: Wq
Please enter a valid integer for maximum mismatches.
Enter maximum number of mismatches TATA box exist/default=0:
```

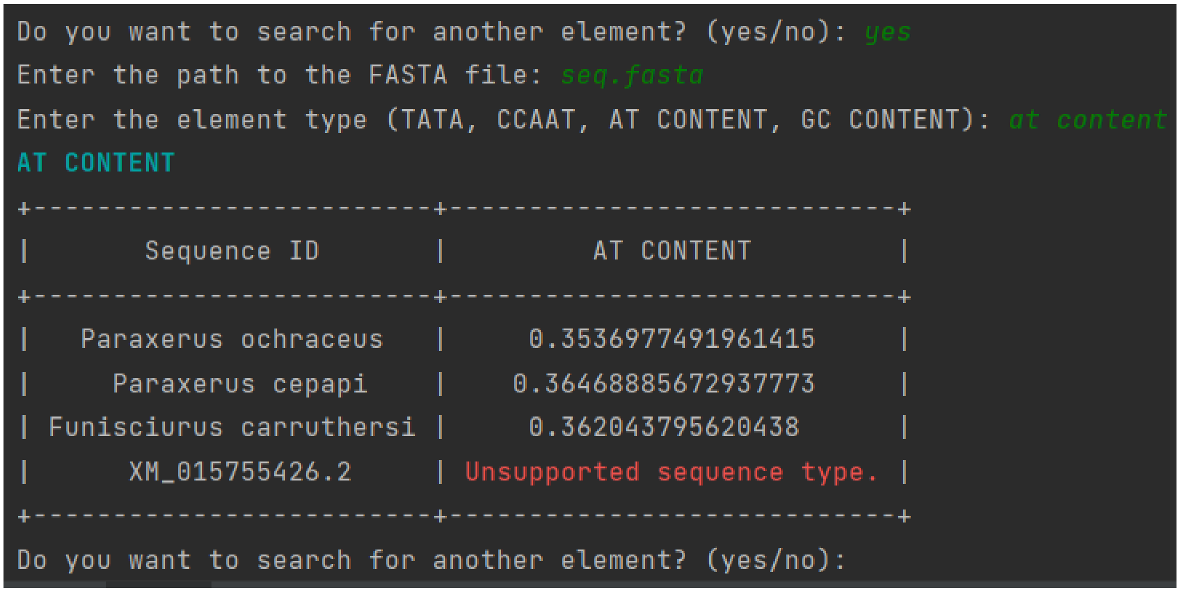
CCAAT Box



```
Do you want to search for another element? (yes/no): yes
Enter the path to the FASTA file: seq.fasta
Enter the element type (TATA, CCAAT, AT CONTENT, GC CONTENT): ccaat
Enter maximum number of mismatches CCAAT box exist/default=0: 2
CCAAT BOX MOTIF
+-------------------------+----------+--------------------+
|       Sequence ID       | Position |       Motif        |
+-------------------------+----------+--------------------+
|   Paraxerus ochraceus   |   343    |      TGCCAAG       |
|   Paraxerus ochraceus   |   532    |      GCCCAAG       |
|     Paraxerus cepapi    |   398    |      TGCCAAG       |
|     Paraxerus cepapi    |   587    |      GCCCAAG       |
|     Paraxerus cepapi    |   671    |      GGCCAGC       |
| Funisciurus carruthersi |   395    |      TGCCAAG       |
| Funisciurus carruthersi |   584    |      GCCCAAG       |
| Funisciurus carruthersi |   668    |      GGCCAGC       |
|      XM_015755426.2     |   N/A    | Not a DNA sequence |
+-------------------------+----------+--------------------+
Do you want to search for another element? (yes/no):
```

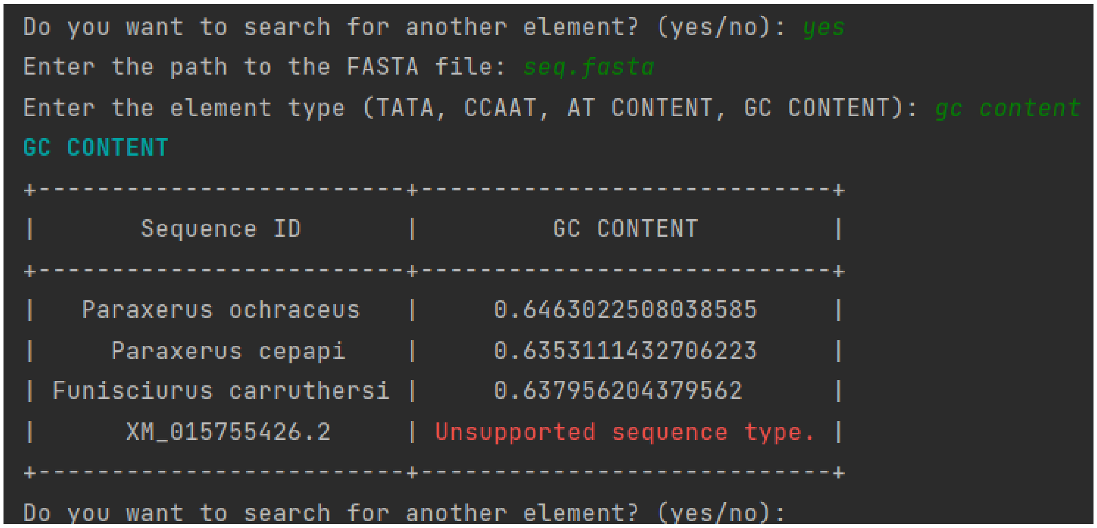
Index - 15743

AT Content



```
Do you want to search for another element? (yes/no): yes
Enter the path to the FASTA file: seq.fasta
Enter the element type (TATA, CCAAT, AT CONTENT, GC CONTENT): at content
AT CONTENT
+-------------------------+----------------------------+
|       Sequence ID       |         AT CONTENT         |
+-------------------------+----------------------------+
|   Paraxerus ochraceus   |     0.3536977491961415     |
|     Paraxerus cepapi    |    0.36468885672937773     |
| Funisciurus carruthersi |     0.362043795620438      |
|      XM_015755426.2     | Unsupported sequence type. |
+-------------------------+----------------------------+
Do you want to search for another element? (yes/no):
```

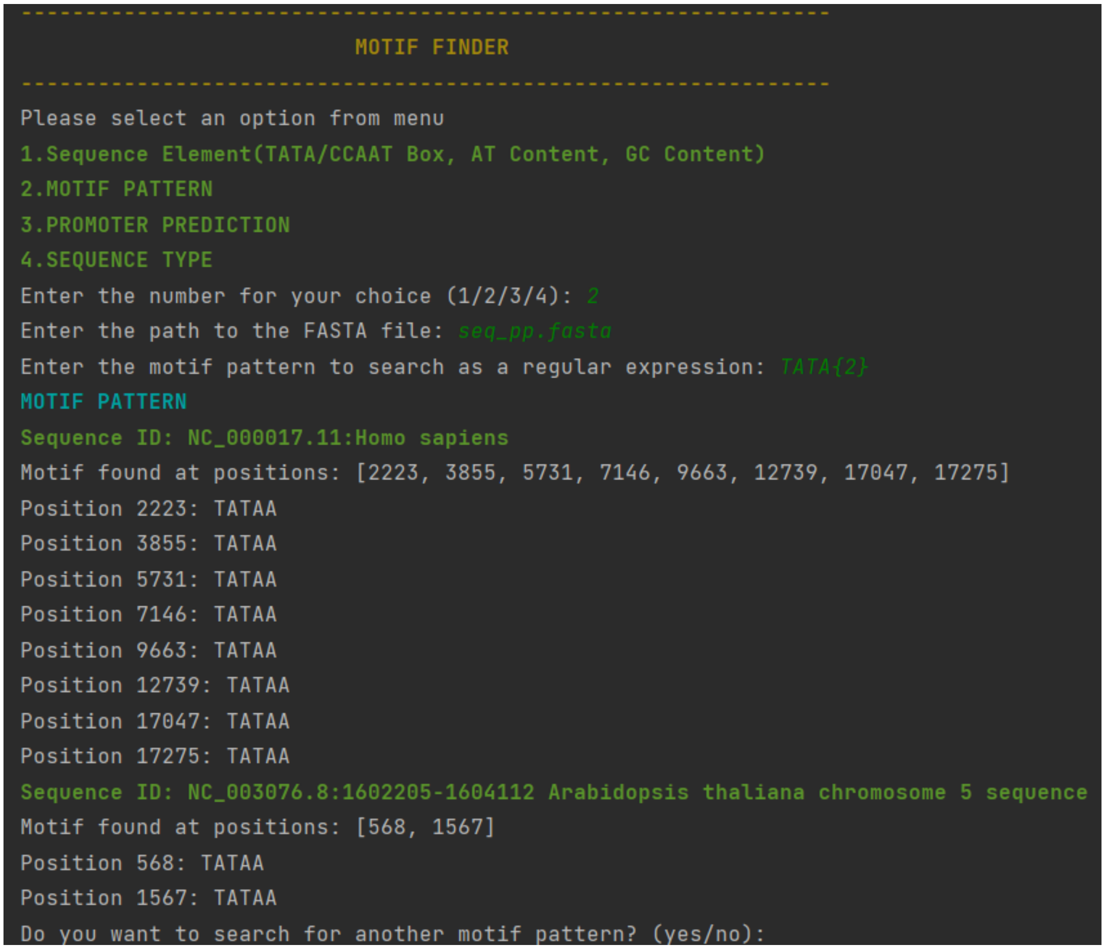
GC Content



```
Do you want to search for another element? (yes/no): yes
Enter the path to the FASTA file: seq.fasta
Enter the element type (TATA, CCAAT, AT CONTENT, GC CONTENT): gc content
GC CONTENT
+-------------------------+----------------------------+
|       Sequence ID       |         GC CONTENT         |
+-------------------------+----------------------------+
|   Paraxerus ochraceus   |     0.6463022508038585     |
|     Paraxerus cepapi    |     0.6353111432706223     |
| Funisciurus carruthersi |     0.637956204379562      |
|      XM_015755426.2     | Unsupported sequence type. |
+-------------------------+----------------------------+
Do you want to search for another element? (yes/no):
```

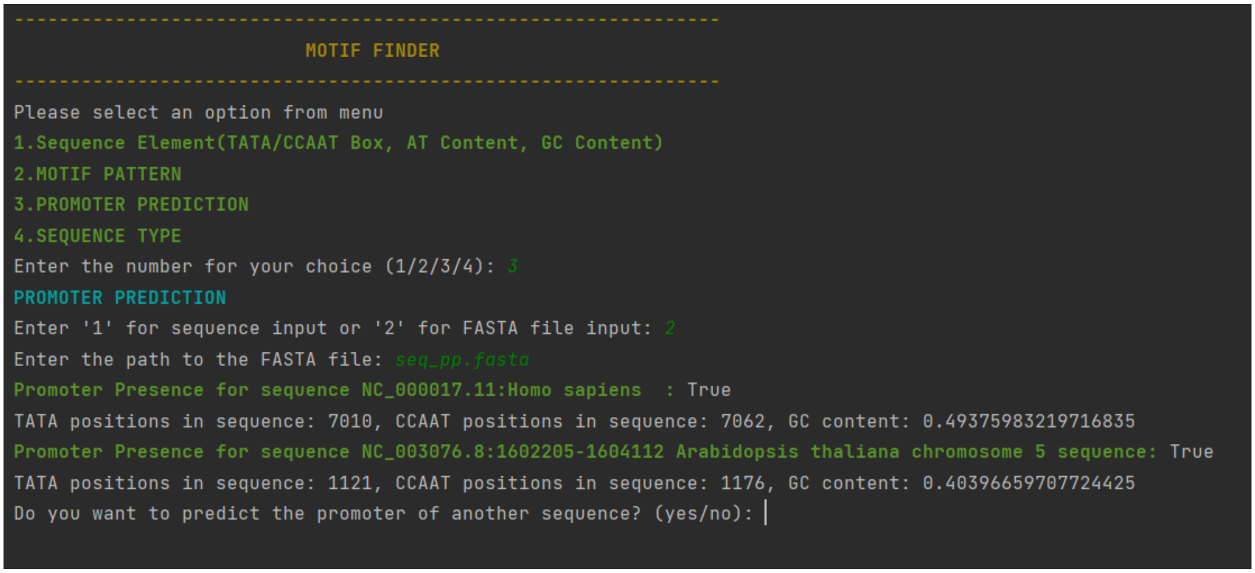
Index - 15743

Motif Pattern



```
---------------------------------------------------------------
                          MOTIF FINDER
---------------------------------------------------------------
Please select an option from menu
1.Sequence Element(TATA/CCAAT Box, AT Content, GC Content)
2.MOTIF PATTERN
3.PROMOTER PREDICTION
4.SEQUENCE TYPE
Enter the number for your choice (1/2/3/4): 2
Enter the path to the FASTA file: seq_pp.fasta
Enter the motif pattern to search as a regular expression: TATA{2}
MOTIF PATTERN
Sequence ID: NC_000017.11:Homo sapiens
Motif found at positions: [2223, 3855, 5731, 7146, 9663, 12739, 17047, 17275]
Position 2223: TATAA
Position 3855: TATAA
Position 5731: TATAA
Position 7146: TATAA
Position 9663: TATAA
Position 12739: TATAA
Position 17047: TATAA
Position 17275: TATAA
Sequence ID: NC_003076.8:1602205-1604112 Arabidopsis thaliana chromosome 5 sequence
Motif found at positions: [568, 1567]
Position 568: TATAA
Position 1567: TATAA
Do you want to search for another motif pattern? (yes/no):
```

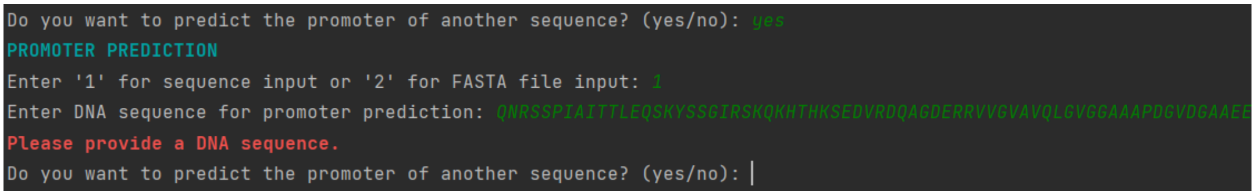
Promoter Prediction



```
---------------------------------------------------------------
                          MOTIF FINDER
---------------------------------------------------------------
Please select an option from menu
1.Sequence Element(TATA/CCAAT Box, AT Content, GC Content)
2.MOTIF PATTERN
3.PROMOTER PREDICTION
4.SEQUENCE TYPE
Enter the number for your choice (1/2/3/4): 3
PROMOTER PREDICTION
Enter '1' for sequence input or '2' for FASTA file input: 2
Enter the path to the FASTA file: seq_pp.fasta
Promoter Presence for sequence NC_000017.11:Homo sapiens  : True
TATA positions in sequence: 7010, CCAAT positions in sequence: 7062, GC content: 0.49375983219716835
Promoter Presence for sequence NC_003076.8:1602205-1604112 Arabidopsis thaliana chromosome 5 sequence: True
TATA positions in sequence: 1121, CCAAT positions in sequence: 1176, GC content: 0.40396659707724425
Do you want to predict the promoter of another sequence? (yes/no):
```

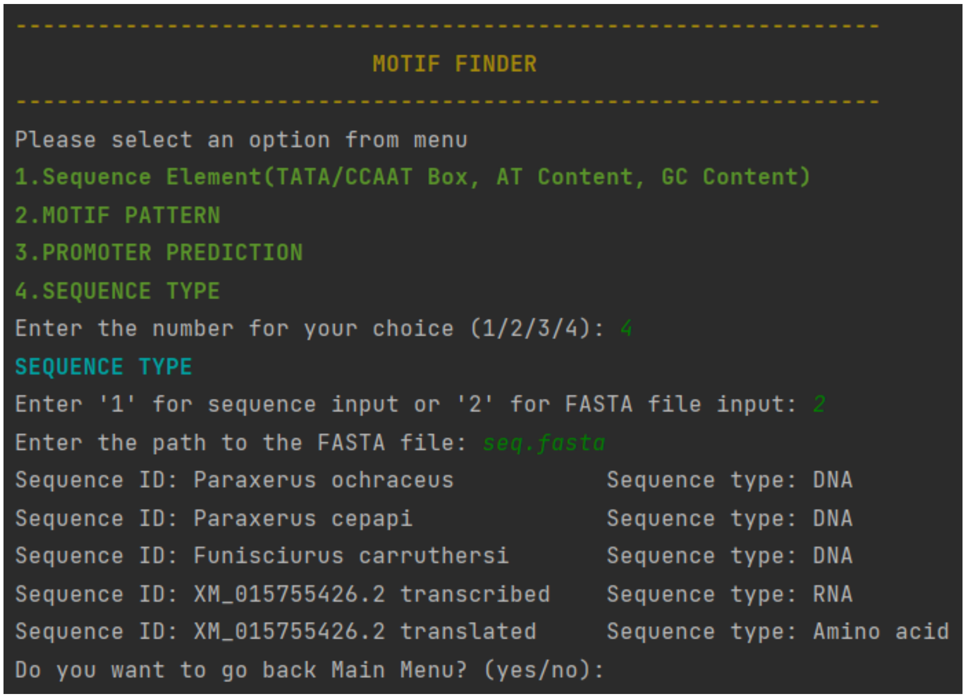
Index - 15743

Provide mRNA sequence



```
Do you want to predict the promoter of another sequence? (yes/no): yes
PROMOTER PREDICTION
Enter '1' for sequence input or '2' for FASTA file input: 1
Enter DNA sequence for promoter prediction: QNRSSPIAITTLEQSKYSSGIRSKQKHTHKSEDVRDQAGDERRVVGVAVQLGVGGAAAPDGVDGAAEE
Please provide a DNA sequence.
Do you want to predict the promoter of another sequence? (yes/no):
```

Sequence Type



```
---------------------------------------------------------------
                          MOTIF FINDER
---------------------------------------------------------------
Please select an option from menu
1.Sequence Element(TATA/CCAAT Box, AT Content, GC Content)
2.MOTIF PATTERN
3.PROMOTER PREDICTION
4.SEQUENCE TYPE
Enter the number for your choice (1/2/3/4): 4
SEQUENCE TYPE
Enter '1' for sequence input or '2' for FASTA file input: 2
Enter the path to the FASTA file: seq.fasta
Sequence ID: Paraxerus ochraceus           Sequence type: DNA
Sequence ID: Paraxerus cepapi              Sequence type: DNA
Sequence ID: Funisciurus carruthersi       Sequence type: DNA
Sequence ID: XM_015755426.2 transcribed    Sequence type: RNA
Sequence ID: XM_015755426.2 translated     Sequence type: Amino acid
Do you want to go back Main Menu? (yes/no):
```

References

1. Implementation of dRNA-seq-driven, species-specific promoter prediction using convolutional neural networks, Authors:Lucas Coppens, Laura Wicke, KU Leuven, Rob Lavigne, KU Leuven https://www.researchgate.net/publication/363411219_SAPPHIRECNN_Implementation_of_dRNA-seq-driven_species-specific_promoter_prediction_using_convolutional_neural_networks

2. Identification of putative promoters in 48 eukaryotic genomes on the basis of DNA free energy, Venkata Rajesh Yella, Aditya Kumar and Manju Bansal, PMID: 29540741

3. de Avila e Silva S, Echeverrigaray S, Gerhardt GJL. BacPP: Bacterial promoterprediction-A tool for accurate sigma-factor specific assignment inenterobacteria. J Theor Biol 2011;287(1):92–9. https://doi.org/10.1016/j.jtbi.2011.07.017.

Index - 15743