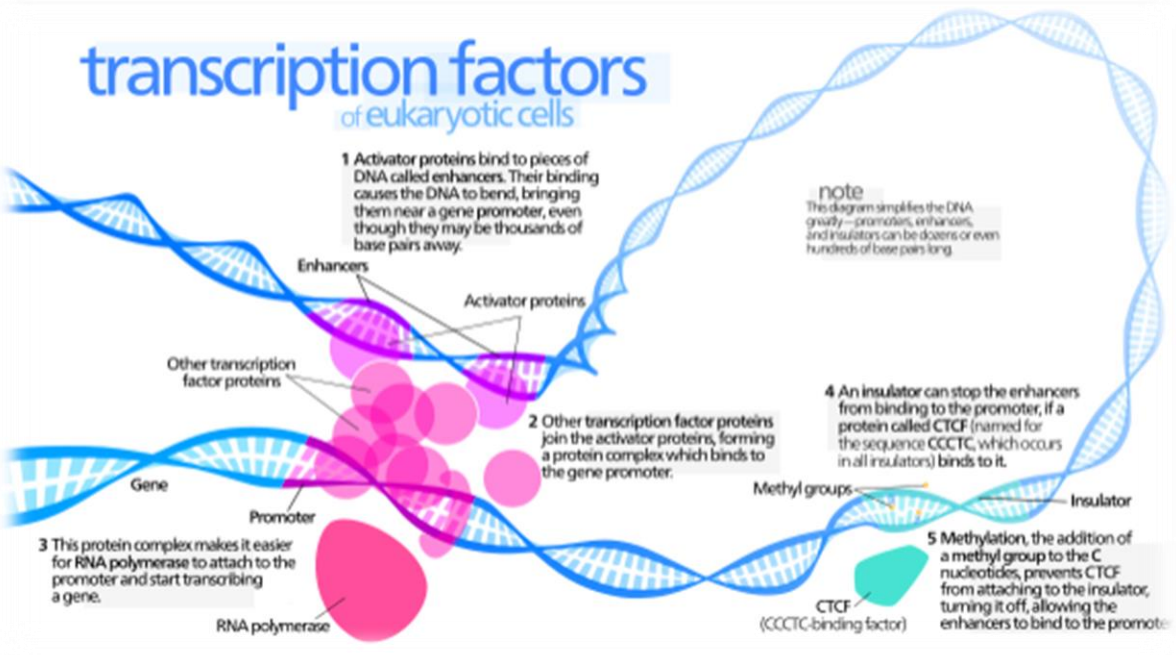
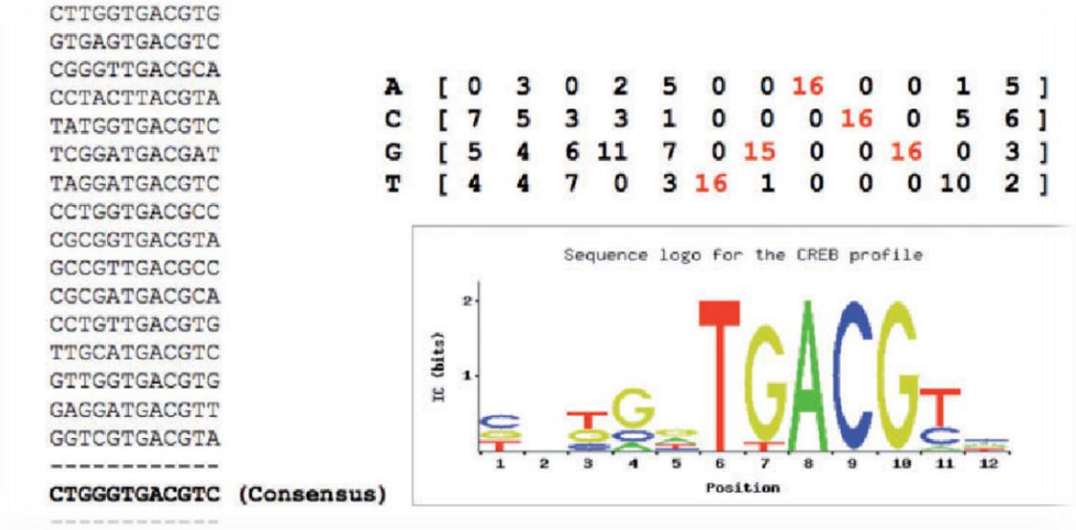
A tool to discover motifs among DNA sequences using Expectation Maximization (EM) algorithm





Huimin Lu

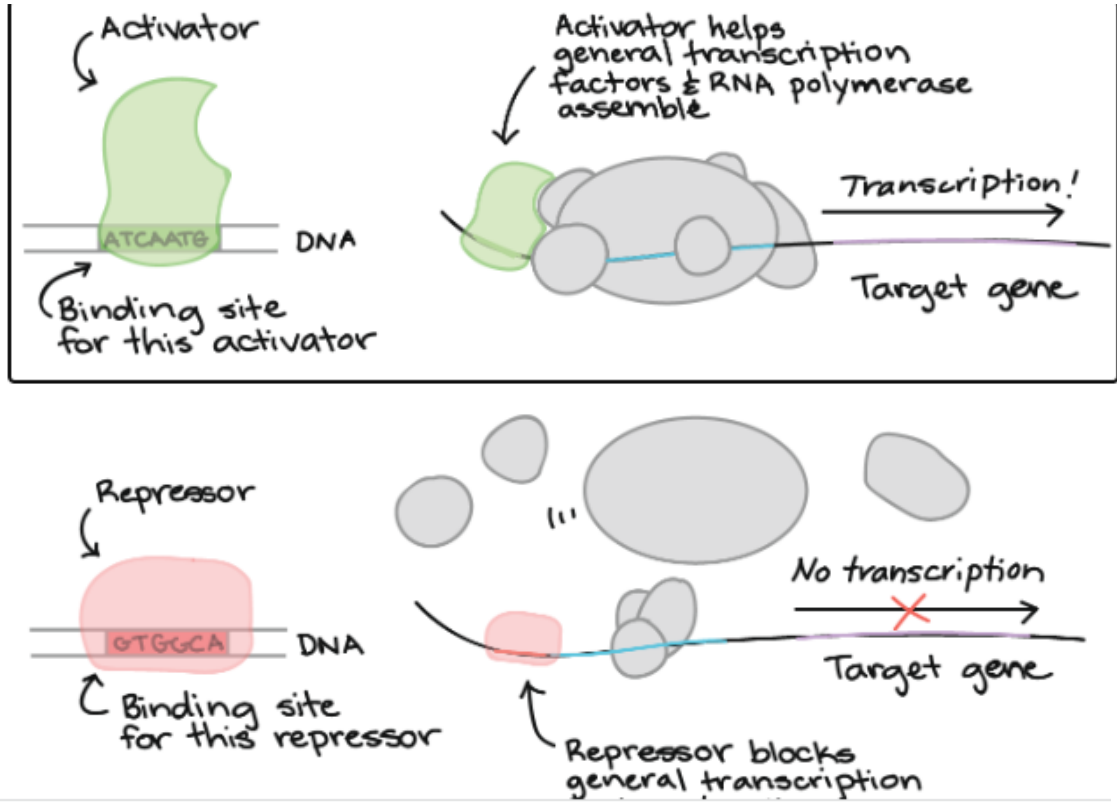
2023-03-14

Background

- In transctription, **transcription facters binding sites** are short seqeunces that combine transcription factor (TF).
- Transtriftion factors could activate or deacticate gene transctiption
- And transcription facters binding sites are highly conserved.
- So, discovering TF binding sites sequence, motif,

  could help find **potential drug target** and genetic

  desease diagnosis.
- And this tool is to find motifs among DNA sequences



Tool presentation

- Three groups of **input dataset**
- One **python scipt** and import other python softeware called “Weblogo”.
- Finally getting a **motif logo**,

Package of this project

conda	22.9.0
weblogo	3.7
python	3.9
pandas	1.4
numpy	1.21

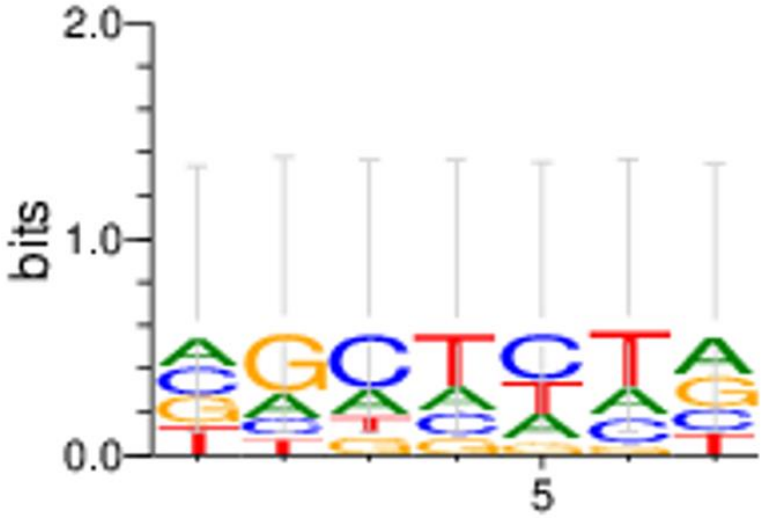
Python program

```
usage: main.py [-h] [-kmer KMER] [-A A] [-T T] [-C C] [-G G] [-X X] input_file output_file

Replication EM alorgrithm in DNA motif discovery

positional arguments:
  input_file   input fasta file, it must be standard fasta file
  output_file  output txt file, this file will be used to make motif logo

options:
  -h, --help   show this help message and exit
  -kmer KMER   the length of motif
  -A A         background(non motif probability)
  -T T         background(non motif probability)
  -C C         background(non motif probability)
  -G G         background(non motif)probability
  -X X         probalility of motif among dataset, used in calculate initial p matrix
```

Input file

```
>seq1
CTATAAACGTTACA
>seq2
ATAGCGATTCGACT
>seq3
CAGCCCAGAACCCT
>seq4
CGGTGAACCTTACA
>seq5
TGCATTCAATAGCT
>seq6
TGCTCTGTCCACTC
>seq7
TGCTCTGTCCACTC
>seq8
GGTCTACCTTTATC
```

output



Significance and novelty

Complexilarity:

- This tool does not use regular expression, instead using Expectstion Maximization Algorithm, **iterating update**.
- This helps to find more potention motif whose patterns are difficult to predict using Regtex.
- The most difficult part is **to understand this algorithm and compute by my own script**

Feature:

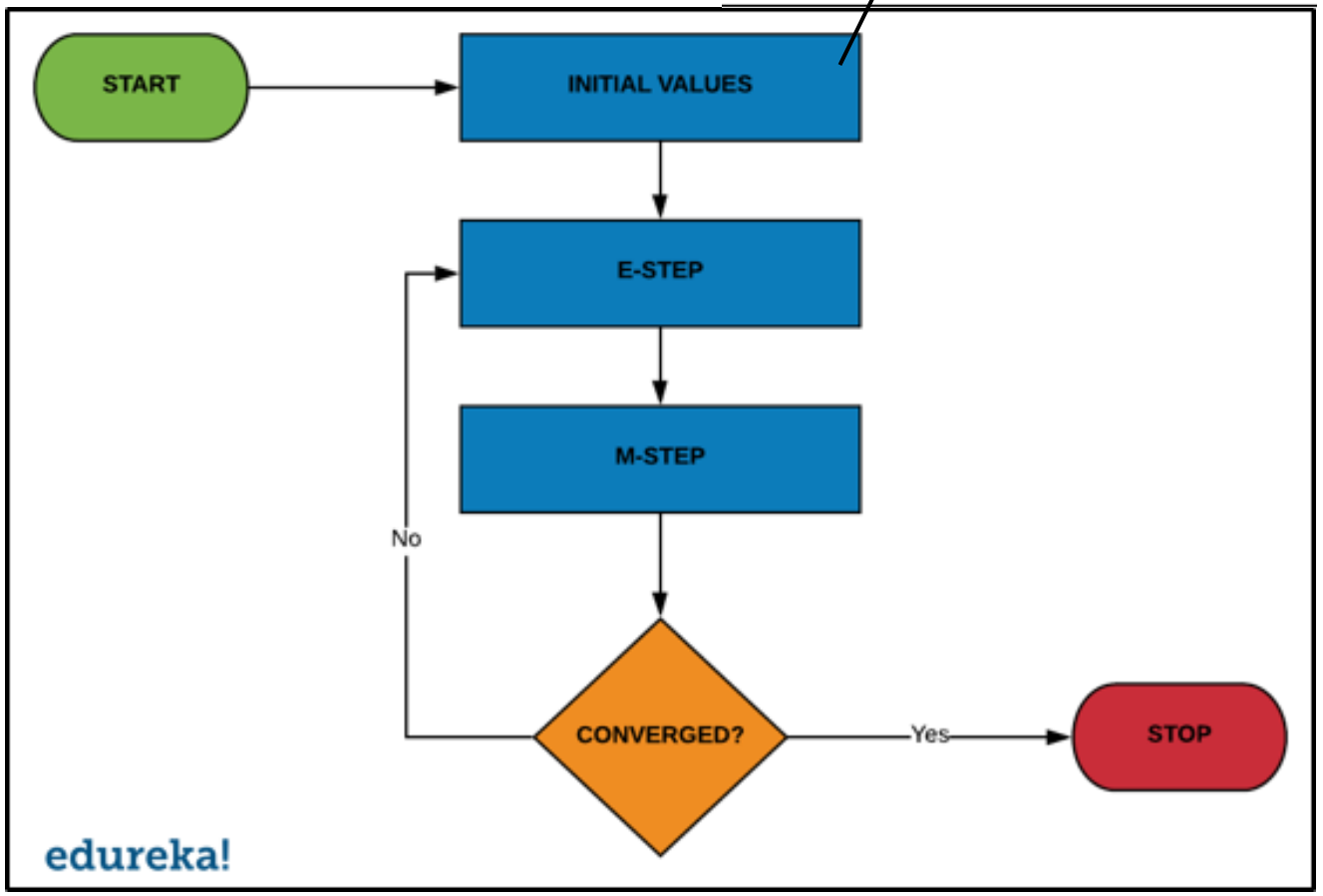
- This tool can calculate motif profile matrix and make motif logo

Limitation:

- Only discover DNA sequences nor protein sequences.

Methodology

$$p = \begin{array}{c|ccc} & 1 & 2 & 3 \\ A & 0.17 & 0.5 & 0.17 \\ C & 0.17 & 0.17 & 0.17 \\ G & 0.17 & 0.17 & 0.17 \\ T & 0.5 & 0.17 & 0.5 \end{array}$$



- **E-step**

Example: Estimating Z

$X_i =$ G C T G T A G

$$p = \begin{array}{c|cccc} & 0 & 1 & 2 & 3 \\ A & 0.25 & 0.1 & 0.5 & 0.2 \\ C & 0.25 & 0.4 & 0.2 & 0.1 \\ G & 0.25 & 0.3 & 0.1 & 0.6 \\ T & 0.25 & 0.2 & 0.2 & 0.1 \end{array}$$

$$Z_{i1} = 0.3 \times 0.2 \times 0.1 \times 0.25 \times 0.25 \times 0.25 \times 0.25$$

$$Z_{i2} = 0.25 \times 0.4 \times 0.2 \times 0.6 \times 0.25 \times 0.25 \times 0.25$$

$\vdots$

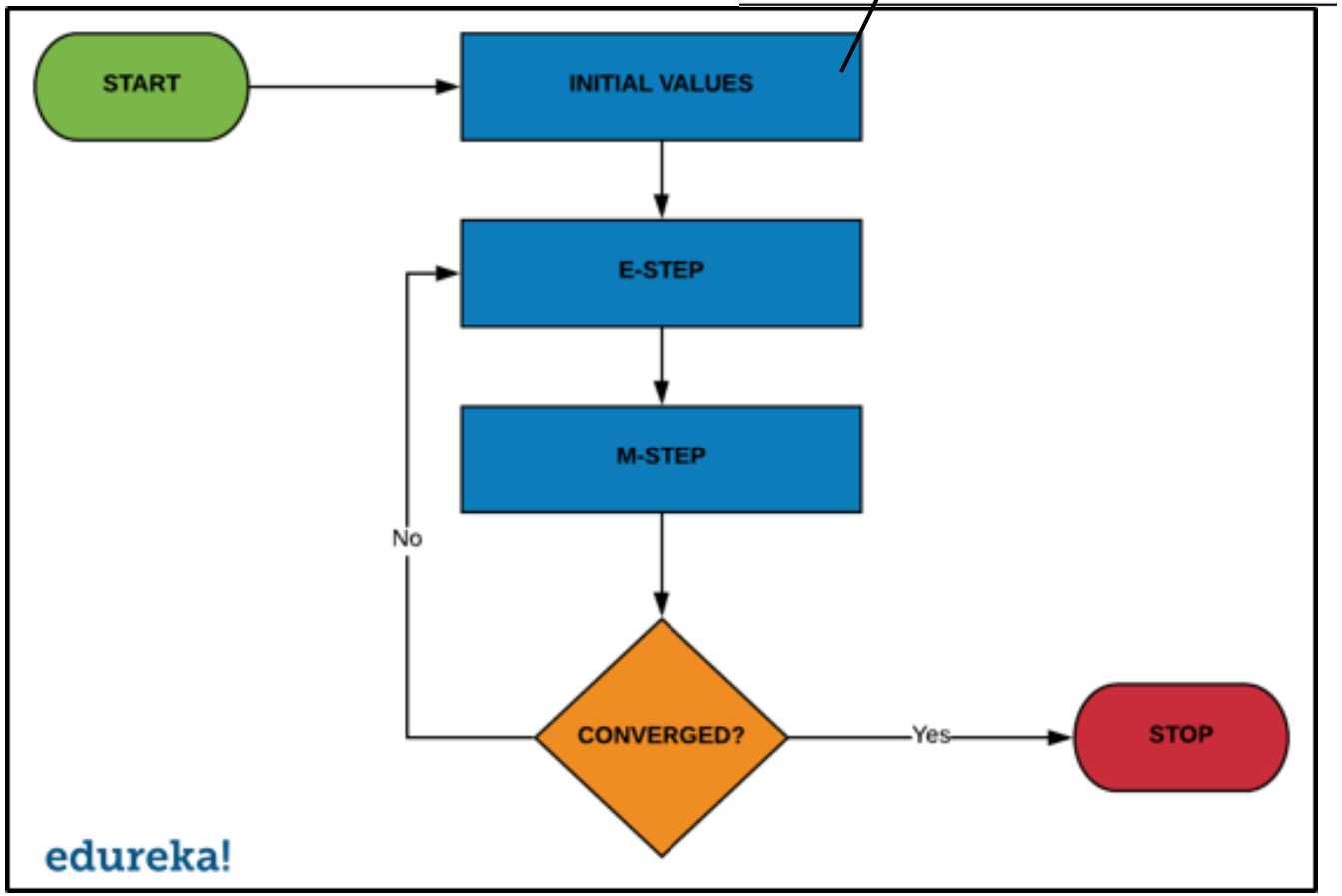
- then normalize so that $\sum_{j=1}^{L-W+1} Z_{ij} = 1$

$$Z = \begin{array}{c|cccc} & 1 & 2 & 3 & 4 \\ seq1 & 0.1 & 0.1 & 0.2 & 0.6 \\ seq2 & 0.4 & 0.2 & 0.1 & 0.3 \\ seq3 & 0.3 & 0.1 & 0.5 & 0.1 \\ seq4 & 0.1 & 0.5 & 0.1 & 0.3 \end{array}$$

Methodology

$p =$

	1	2	3
A	0.17	0.5	0.17
C	0.17	0.17	0.17
G	0.17	0.17	0.17
T	0.5	0.17	0.5



- **M-step**

The M-step: Estimating p

- recall $p_{c,k}$ represents the probability of character c in position k ; values for position 0 represent the background

$$p_{c,k}^{(t+1)} = \frac{n_{c,k} + d_{c,k}}{\sum_b (n_{b,k} + d_{b,k})}$$

pseudo-counts

$$n_{c,k} = \begin{cases} \sum_i \sum_{\{j | X_{i,j+k-1} = c\}} Z_{ij} & k > 0 \\ n_c - \sum_{j=1}^{W} n_{c,j} & k = 0 \end{cases}$$

total # of c's in data set

Example: Estimating p

A C A G C A

$Z_{1,1} = 0.1, Z_{1,2} = 0.7, Z_{1,3} = 0.1, Z_{1,4} = 0.1$

A G G C A G

$Z_{2,1} = 0.4, Z_{2,2} = 0.1, Z_{2,3} = 0.1, Z_{2,4} = 0.4$

T C A G T C

$Z_{3,1} = 0.2, Z_{3,2} = 0.6, Z_{3,3} = 0.1, Z_{3,4} = 0.1$

$$p_{A,1} = \frac{Z_{1,1} + Z_{1,3} + Z_{2,1} + Z_{3,3} + 1}{Z_{1,1} + Z_{1,2} \ldots + Z_{3,3} + Z_{3,4} + 4}$$

Discuss

- **Pro:**

  First time replicating algorithm

  Learn skils in python programming

- **Cons:**

  Did not find good method to calculating convergence

  The result is not that precise