Motif Search.
Multiple Alignments

Anna Rybina
anna.rybina@skoltech.ru
Bioinformatics course
12.11.2022

inspired by materials from Aleksandra Galitsyna 💛

Skoltech
Skolkovo Institute of Science and Technology

Outline

1. Motif, motif discovery problem ~ 15
2. Multiple sequence alignment ~ 20
3. Task 1 (part of HW) ~ 30
4. ChIP-seq ~ 10
5. Motif representation ~ 20
6. Task 2-4 (part of HW) ~ 45
7. Motif scanning ~ 10
8. In class command line training ~ 10
9. Task 5 (part of HW)

Motif

In general, **motif** is a recurring (**conserved**) **pattern** that is presumed to have **biological significance** (have biological function)

can found in:

- structure or sequence
- RNA/protein/DNA

can be involved in **interactions** with other molecules (proteins/nucleic acids)

Major **living processes** of the cells are **regulated** via **interactions** between proteins and nucleic acids: protein-DNA, protein-RNA, RNA-DNA

During our seminar, we will deal with **DNA sequence motifs** recognized by protein (**transcription factor**)

DNA sequence motif

Often indicate sequence-specific binding sites for proteins such as nucleases and transcription factors (TFs)
Others are involved in processes such as: ribosome binding, mRNA processing (splicing etc), transcription termination

Examples:

- transcription factor binding sites (TFBS)
- motifs recognized by RNA polimerase (e.g. TATA-box in the promoters of *E. coli* genes)
- restriction sites in *E. coli* genome

Motif knowledge is very useful in defining genetic regulatory networks and regulatory program of individual genes

How to find a DNA sequence motif?

Experimental approach:

- DNase footprinting
- SELEX
- electrophoretic mobility shift assays
- more examples https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3080775/

Computational approach – **motif discovery problem**

- search for overrepresented (and/or conserved) DNA patterns upstream of functionally related genes (e.g. genes with similar expression patterns)

Motif discovery problem

It is the computational task of searching for regulatory DNA motifs

The motif discovery problem can be formulated as follows:

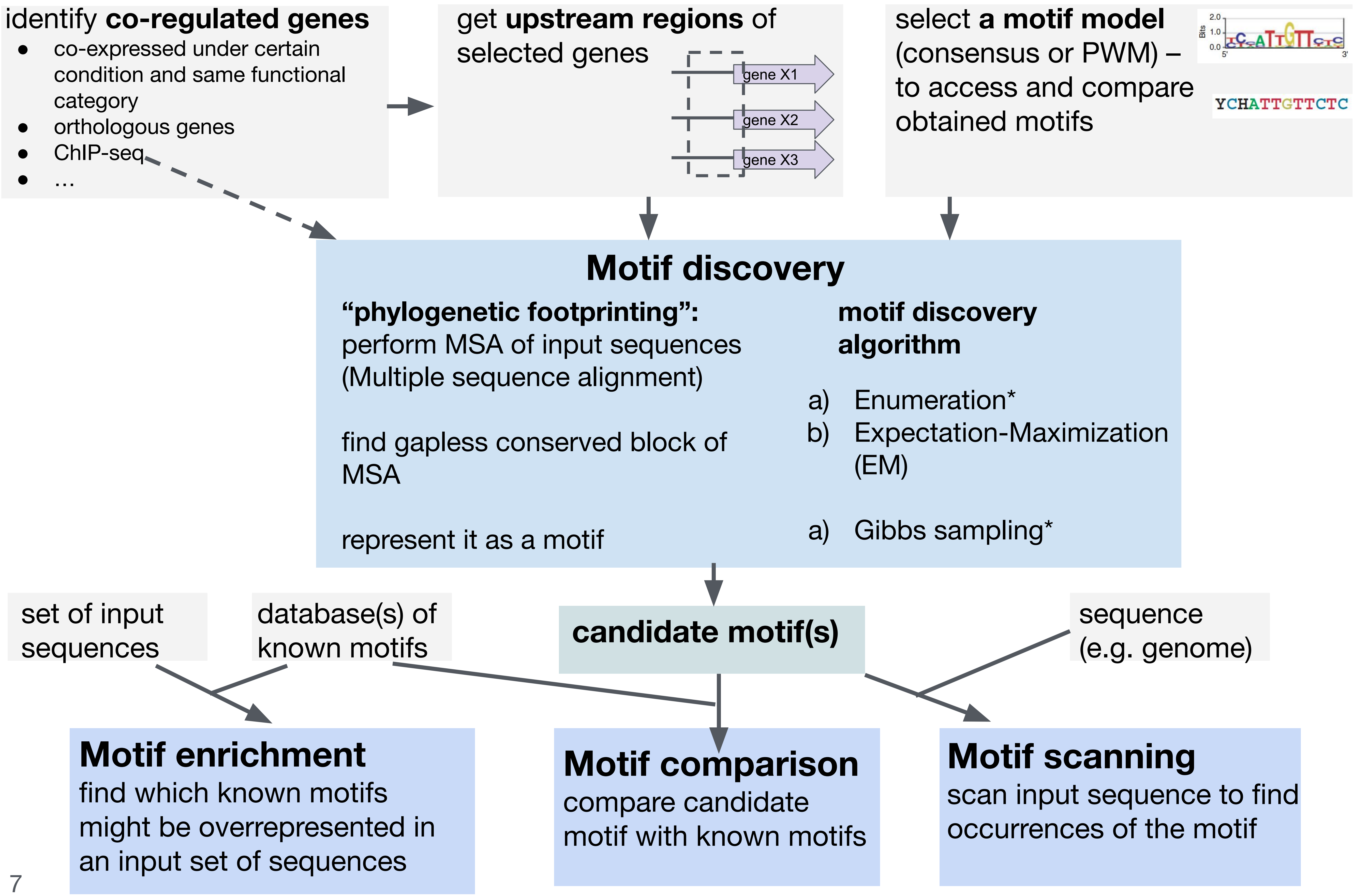
Given: a set of DNA sequences

Assumption: respective genes are **co-regulated** and thus likely to be bound by one or more regulatory proteins

Find: parameters of **motif(s)** that could explain this binding:

- number of motifs
- the width of each motif
- its location in input sequences

Motif discovery problem: flowchart



identify **co-regulated genes**

- co-expressed under certain condition and same functional category
- orthologous genes
- ChIP-seq
- …

get **upstream regions** of selected genes

select **a motif model** (consensus or PWM) – to access and compare obtained motifs

Motif discovery

“phylogenetic footprinting”: perform MSA of input sequences (Multiple sequence alignment)

find gapless conserved block of MSA

represent it as a motif

motif discovery algorithm

a) Enumeration*
b) Expectation-Maximization (EM)

a) Gibbs sampling*

set of input sequences

database(s) of known motifs

candidate motif(s)

sequence (e.g. genome)

Motif enrichment
find which known motifs might be overrepresented in an input set of sequences

Motif comparison
compare candidate motif with known motifs

Motif scanning
scan input sequence to find occurrences of the motif

Motif discovery problem: flowchart

identify **co-regulated genes**

- co-expressed under certain condition and same functional category
- orthologous genes
- ChIP-seq
- …

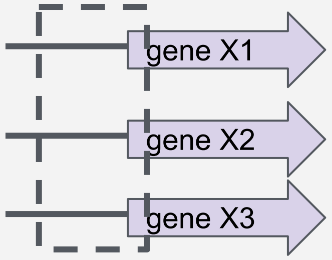
Motif discovery problem: flowchart

identify **co-regulated genes**

- co-expressed under certain condition and same functional category
- orthologous genes
- ChIP-seq
- …

→

get **upstream regions** of selected genes

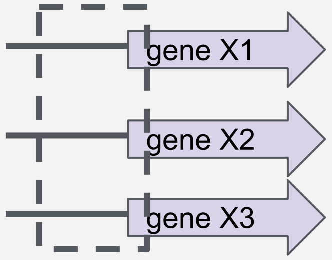


Motif discovery problem: flowchart

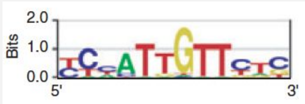
identify **co-regulated genes**

- co-expressed under certain condition and same functional category
- orthologous genes
- ChIP-seq
- …

get **upstream regions** of selected genes



select **a motif model** (consensus or PWM) – to access and compare obtained motifs



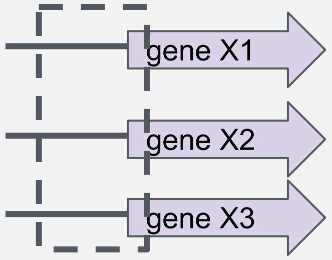
YCHATTGTTCTC

Motif discovery problem: flowchart

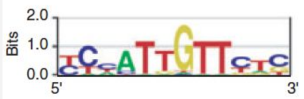
identify **co-regulated genes**

- co-expressed under certain condition and same functional category
- orthologous genes
- ChIP-seq
- …

get **upstream regions** of selected genes



select **a motif model** (consensus or PWM) – to access and compare obtained motifs



YCHATTGTTCTC

Motif discovery

“phylogenetic footprinting”:

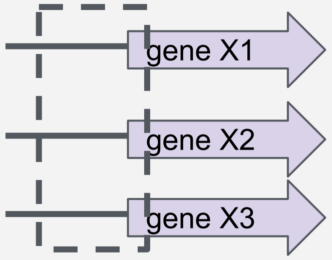
motif discovery algorithm

Motif discovery problem: flowchart

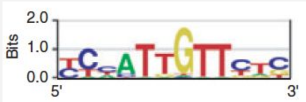
identify **co-regulated genes**

- co-expressed under certain condition and same functional category
- orthologous genes
- ChIP-seq
- …

get **upstream regions** of selected genes



select **a motif model** (consensus or PWM) – to access and compare obtained motifs



YCHATTGTTCTC

Motif discovery

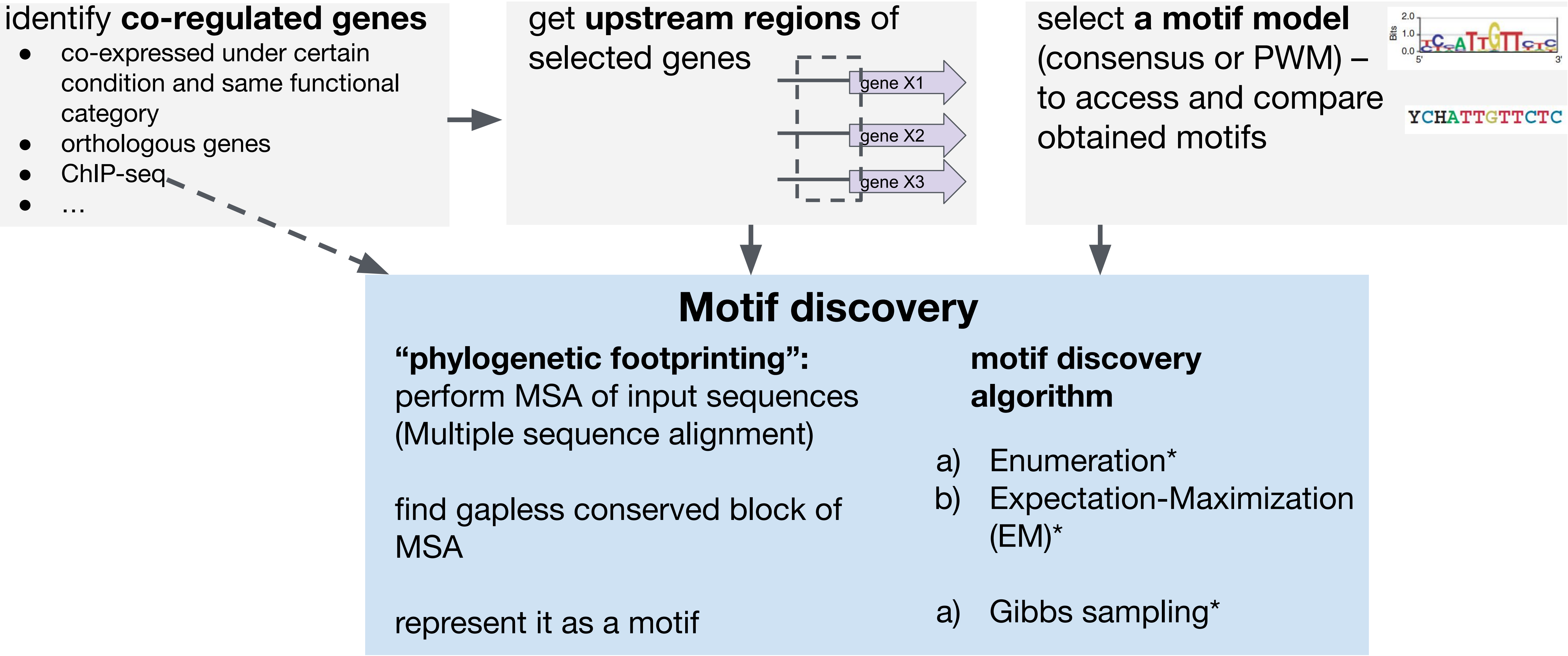
“phylogenetic footprinting”:
perform MSA of input sequences (Multiple sequence alignment)

find gapless conserved block of MSA

represent it as a motif

motif discovery algorithm

Motif discovery problem: flowchart



identify **co-regulated genes**

- co-expressed under certain condition and same functional category
- orthologous genes
- ChIP-seq
- …

get **upstream regions** of selected genes

select **a motif model** (consensus or PWM) – to access and compare obtained motifs

Motif discovery

“phylogenetic footprinting”: perform MSA of input sequences (Multiple sequence alignment)

find gapless conserved block of MSA

represent it as a motif

motif discovery algorithm

a) Enumeration*
b) Expectation-Maximization (EM)*

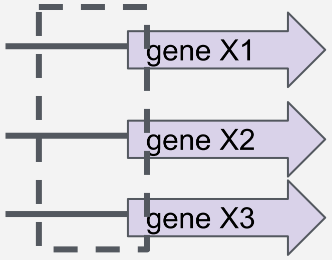
a) Gibbs sampling*

Motif discovery problem: flowchart

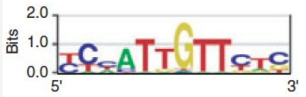
identify **co-regulated genes**

- co-expressed under certain condition and same functional category
- orthologous genes
- ChIP-seq
- …

get **upstream regions** of selected genes



select **a motif model** (consensus or PWM) – to access and compare obtained motifs



YCHATTGTTCTC

Motif discovery

"phylogenetic footprinting":
perform MSA of input sequences (Multiple sequence alignment)

find gapless conserved block of MSA

represent it as a motif

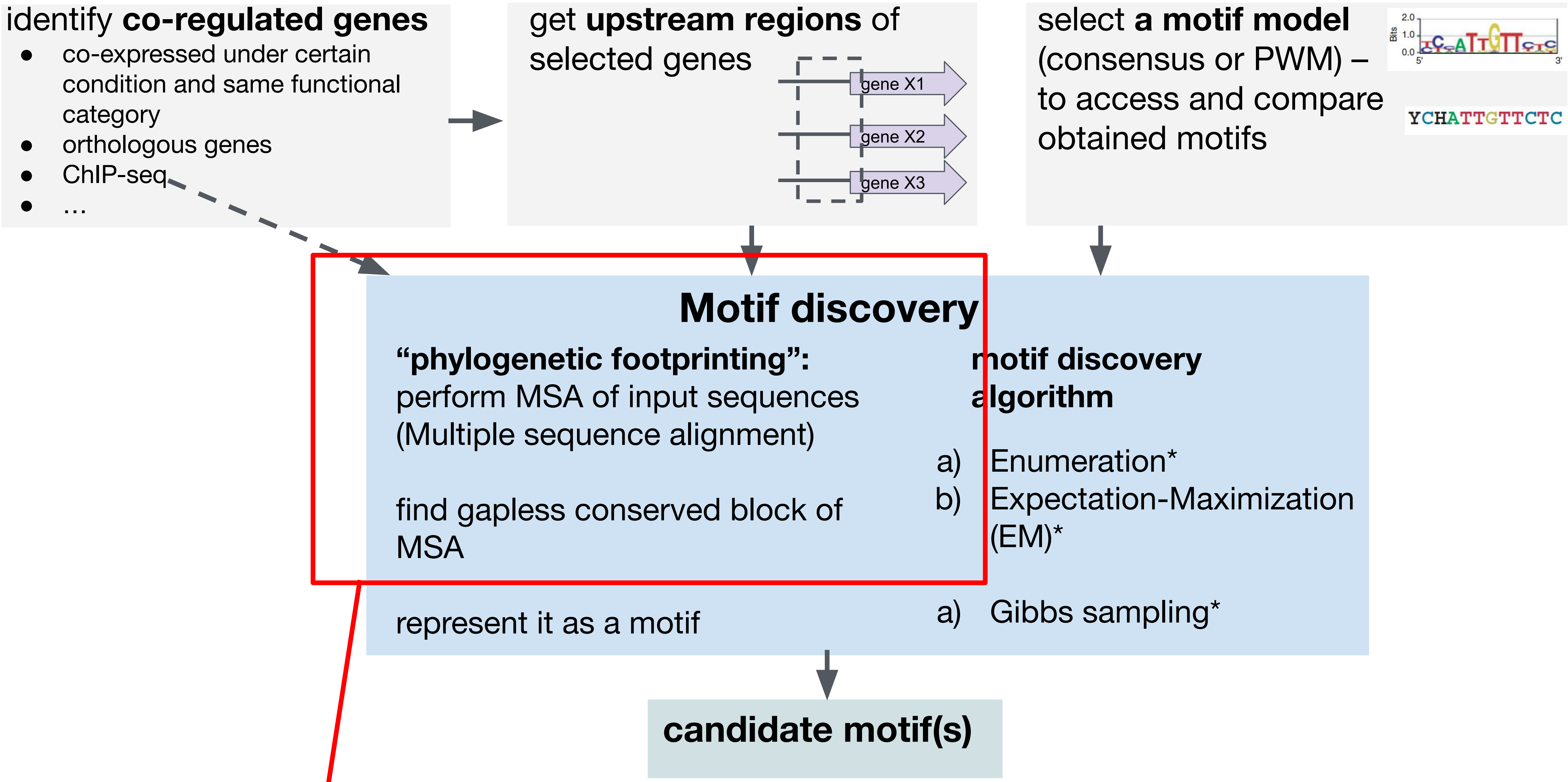
motif discovery algorithm

a) Enumeration*
b) Expectation-Maximization (EM)*

a) Gibbs sampling*

candidate motif(s)

Motif discovery problem: flowchart



identify **co-regulated genes**

- co-expressed under certain condition and same functional category
- orthologous genes
- ChIP-seq
- …

get **upstream regions** of selected genes

select **a motif model** (consensus or PWM) – to access and compare obtained motifs

YCHATTGTTCTC

Motif discovery

“phylogenetic footprinting”: perform MSA of input sequences (Multiple sequence alignment)

find gapless conserved block of MSA

represent it as a motif

motif discovery algorithm

a) Enumeration*
b) Expectation-Maximization (EM)*

a) Gibbs sampling*

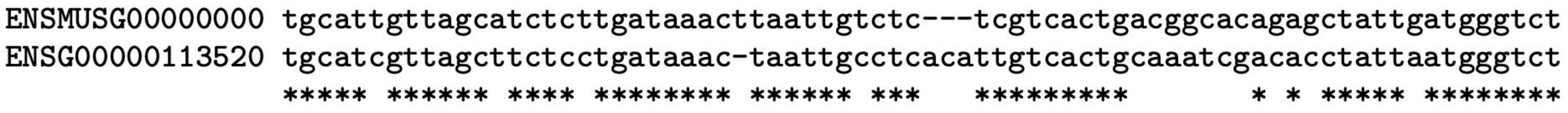
candidate motif(s)

How to use multiple sequence alignment for motif discovery (TASK 1)

Alignment

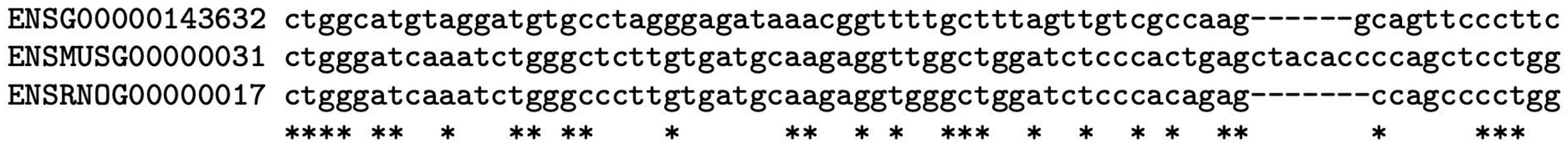
- Can be applied to any sequence (DNA, RNA, protein or other)

- Pairwise alignments (2 sequences):



```
ENSMUSG00000000 tgcattgttagcatctcttgataaacttaattgtctc---tcgtcactgacggcacagagctattgatgggtct
ENSG00000113520 tgcatcgttagcttctcctgataaac-taattgcctcacattgtcactgcaaatcgacacctattaatgggtct
                ***** ****** **** ******** ****** ***   *********       * * ***** ********
```

- Multiple alignments (>=3 sequences):



```
ENSG00000143632 ctggcatgtaggatgtgcctagggagataaacggttttgctttagttgtcgccaag------gcagttcccttc
ENSMUSG00000031 ctgggatcaaatctgggctcttgtgatgcaagaggttggctggatctcccactgagctacaccccagctcctgg
ENSRNOG00000017 ctgggatcaaatctgggcccttgtgatgcaagaggtgggctggatctcccacagag-------ccagcccctgg
                **** **  *   ** **    *      **  * *  ***  *  *  * *  **       *     ***
```

slide courtesy of Asya Mendelevich

Multiple sequence alignment may be used in:

- **phylogenetic analysis**: identify evolutionary relationships between sequences

- **structural bioinformatics**: detect similarities in structure or functions between proteins

- **motif search**: identify shared patterns between sequences

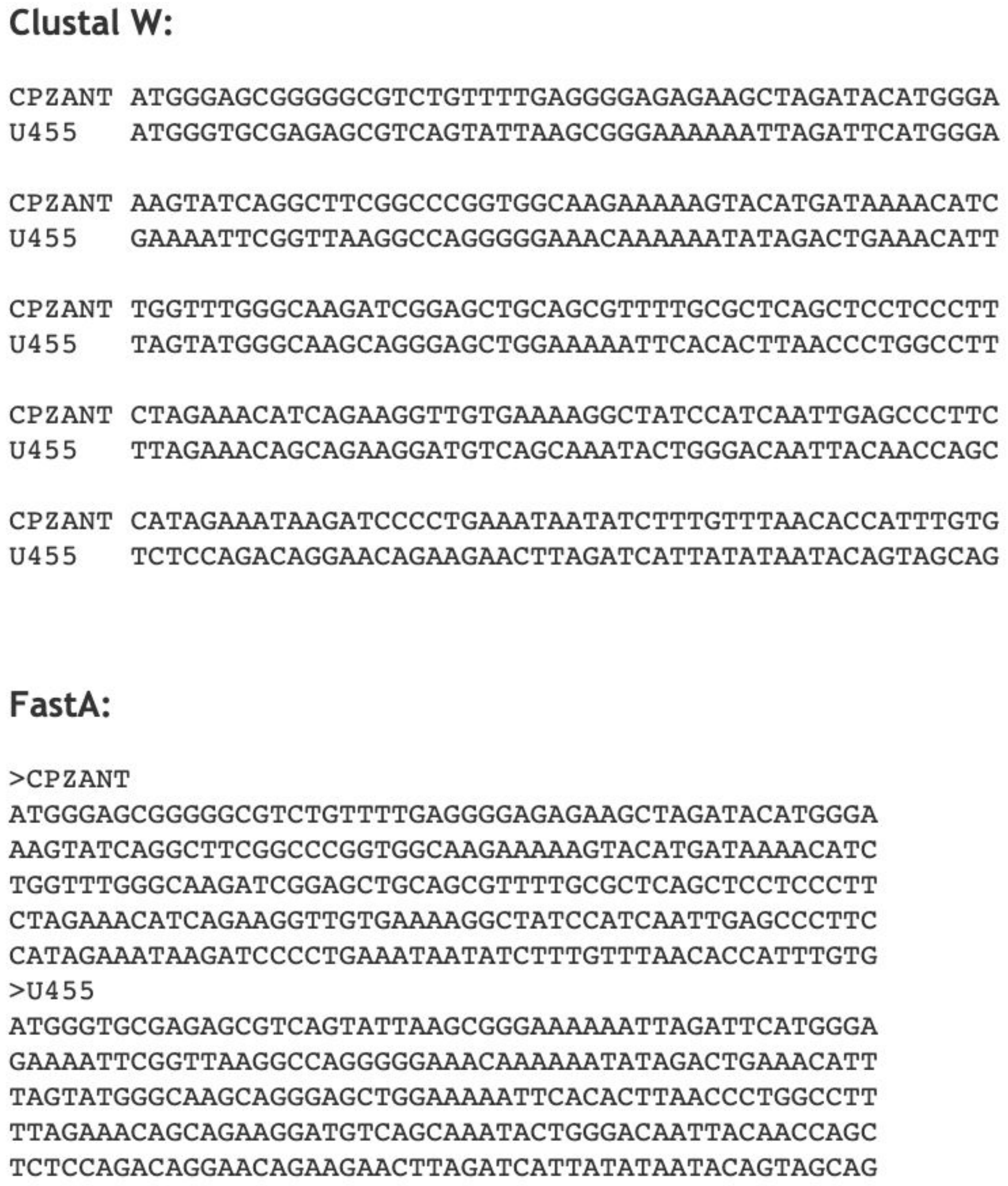
Multiple sequence alignment may be used in:

- **phylogenetic analysis**: identify evolutionary relationships between sequences

- **structural bioinformatics**: detect similarities in structure or functions between proteins

- **motif search**: identify shared patterns between sequences

Alignment formats



Clustal W:

```
CPZANT ATGGGAGCGGGGGCGTCTGTTTTGAGGGGAGAGAAGCTAGATACATGGGA
U455   ATGGGTGCGAGAGCGTCAGTATTAAGCGGGAAAAAATTAGATTCATGGGA

CPZANT AAGTATCAGGCTTCGGCCCGGTGGCAAGAAAAAGTACATGATAAAACATC
U455   GAAAATTCGGTTAAGGCCAGGGGGAAACAAAAAATATAGACTGAAACATT

CPZANT TGGTTTGGGCAAGATCGGAGCTGCAGCGTTTTGCGCTCAGCTCCTCCCTT
U455   TAGTATGGGCAAGCAGGGAGCTGGAAAAATTCACACTTAACCCTGGCCTT

CPZANT CTAGAAACATCAGAAGGTTGTGAAAAGGCTATCCATCAATTGAGCCCTTC
U455   TTAGAAACAGCAGAAGGATGTCAGCAAATACTGGGACAATTACAACCAGC

CPZANT CATAGAAATAAGATCCCCTGAAATAATATCTTTGTTTAACACCATTTGTG
U455   TCTCCAGACAGGAACAGAAGAACTTAGATCATTATATAATACAGTAGCAG
```

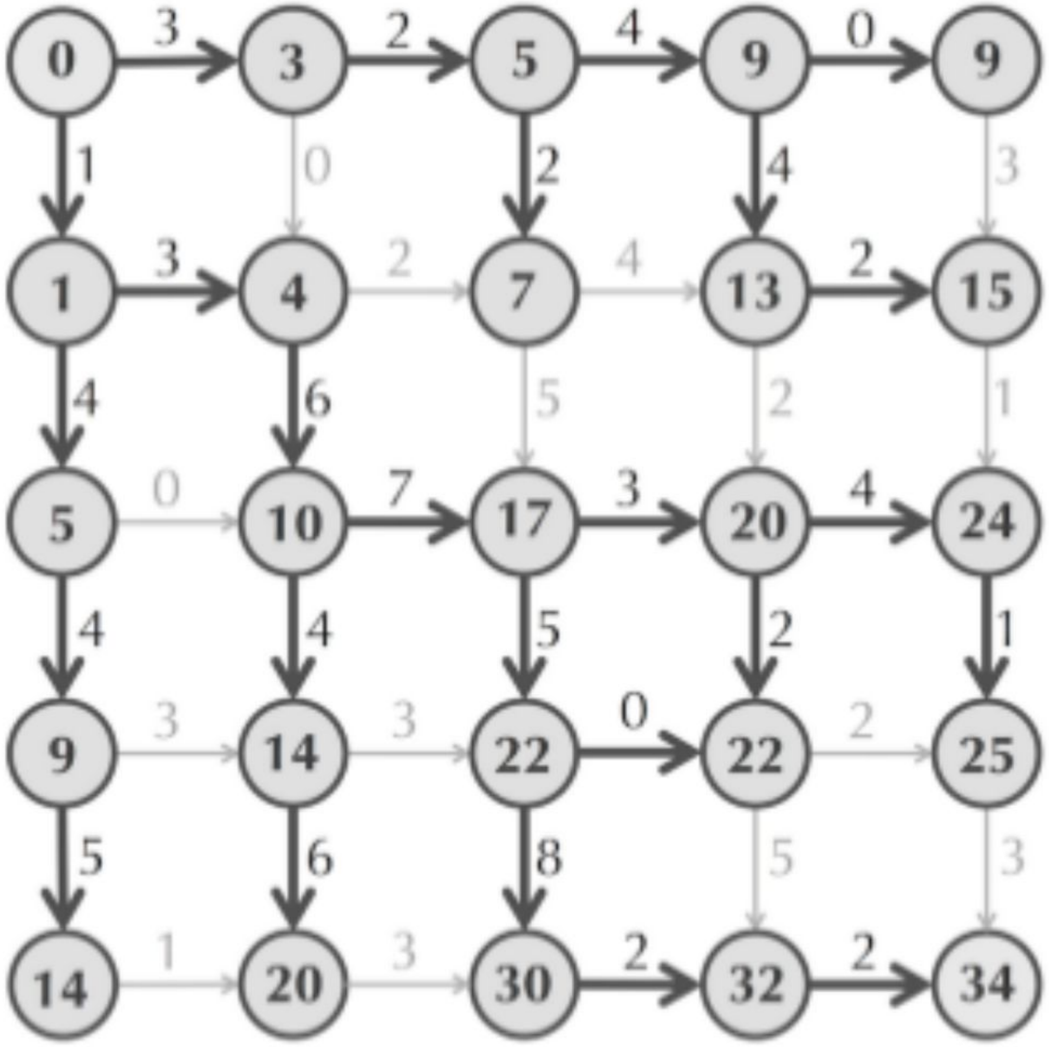
FastA:

```
>CPZANT
ATGGGAGCGGGGGCGTCTGTTTTGAGGGGAGAGAAGCTAGATACATGGGA
AAGTATCAGGCTTCGGCCCGGTGGCAAGAAAAAGTACATGATAAAACATC
TGGTTTGGGCAAGATCGGAGCTGCAGCGTTTTGCGCTCAGCTCCTCCCTT
CTAGAAACATCAGAAGGTTGTGAAAAGGCTATCCATCAATTGAGCCCTTC
CATAGAAATAAGATCCCCTGAAATAATATCTTTGTTTAACACCATTTGTG
>U455
ATGGGTGCGAGAGCGTCAGTATTAAGCGGGAAAAAATTAGATTCATGGGA
GAAAATTCGGTTAAGGCCAGGGGGAAACAAAAAATATAGACTGAAACATT
TAGTATGGGCAAGCAGGGAGCTGGAAAAATTCACACTTAACCCTGGCCTT
TTAGAAACAGCAGAAGGATGTCAGCAAATACTGGGACAATTACAACCAGC
TCTCCAGACAGGAACAGAAGAACTTAGATCATTATATAATACAGTAGCAG
```

More examples:
https://www.hiv.lanl.gov/content/sequence/HelpDocs/SEQsamples.html

Algorithms for multiple sequence alignment

- Ideal: **Dynamic Programming** – optimal solution but not computationally tractable
- **Heuristics*** – approach to reduce the complexity of a problem (~make a computation faster):
    - progressive alignment construction
    - iterative methods
    - consensus methods
    - genetic algorithms

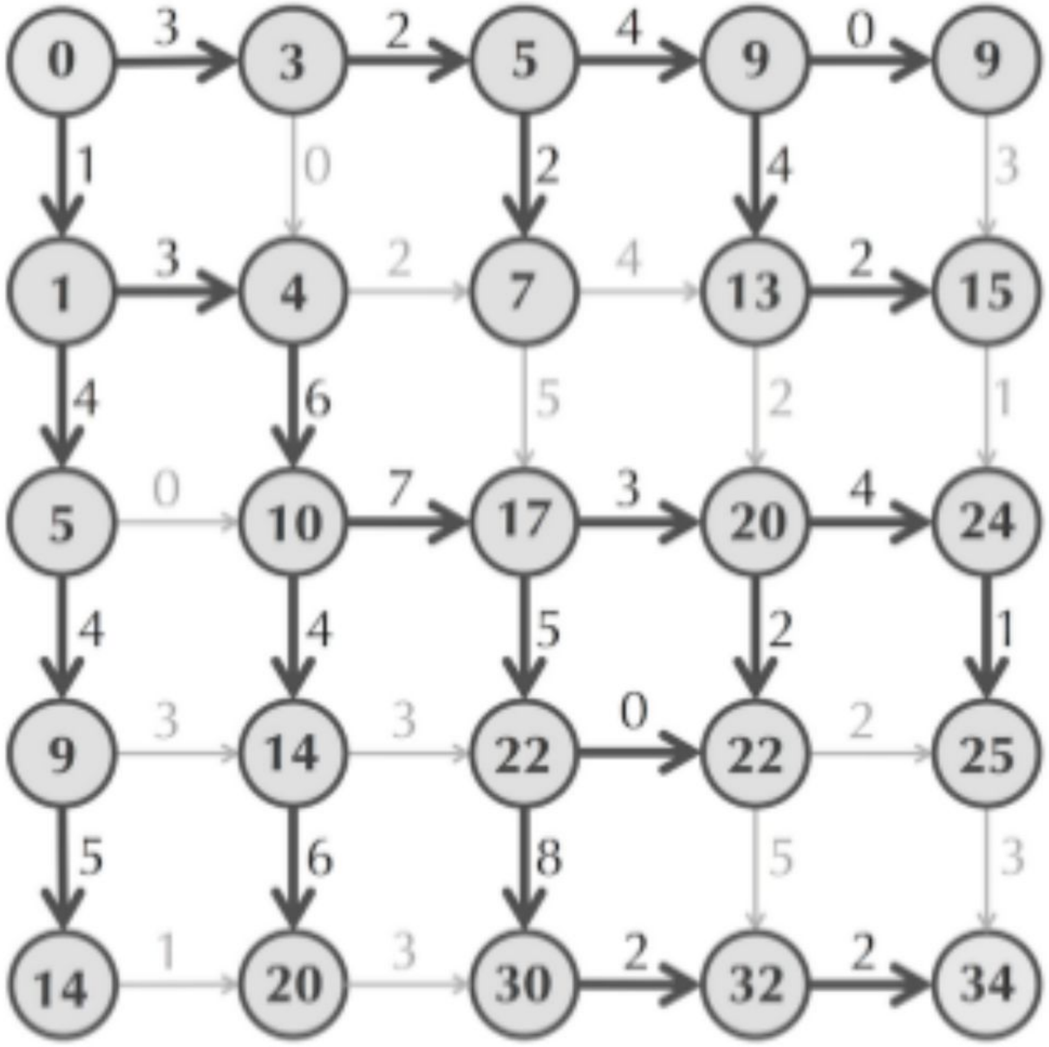


*a ***heuristic*** is an algorithm that is able to ***produce an acceptable solution*** to a problem in many practical scenarios, but for which there is no formal proof of its correctness*

Pavel Pevzner, Phillip Compeau, slide courtesy of Asya Mendelevich

Algorithms for multiple sequence alignment

- Ideal: **Dynamic Programming** – optimal solution but not computationally tractable
- **Heuristics*** – approach to reduce the complexity of a problem (~make a computation faster):
  - progressive alignment construction
  - iterative methods
  - consensus methods
  - genetic algorithms

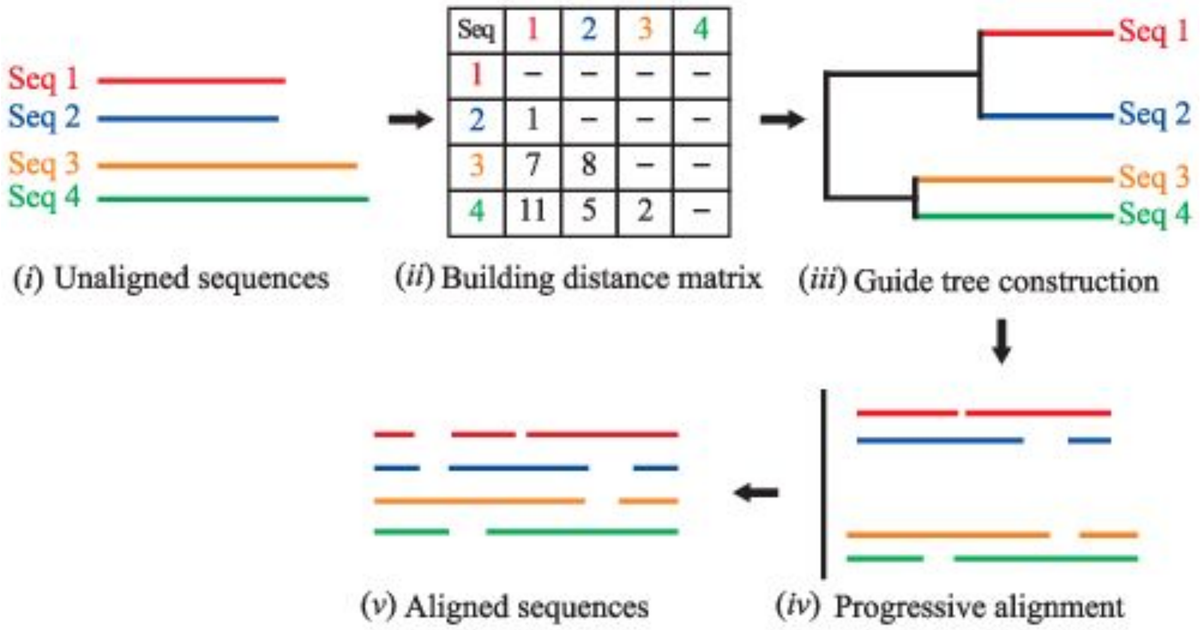


a heuristic* is an algorithm that is able to ***produce an acceptable solution* *to a problem in many practical scenarios, but for which there is no formal proof of its correctness*

Pavel Pevzner, Phillip Compeau, slide courtesy of Asya Mendelevich

Progressive alignments

1. Pairwise alignments (each pair of sequences)
2. Builds a distance matrix
3. Finds a guiding tree using one of clustering methods
4. Builds a multiple alignment progressively, starting from most similar sequences, stacking them as in the guiding tree



Seq	1	2	3	4
1	–	–	–	–
2	1	–	–	–
3	7	8	–	–
4	11	5	2	–

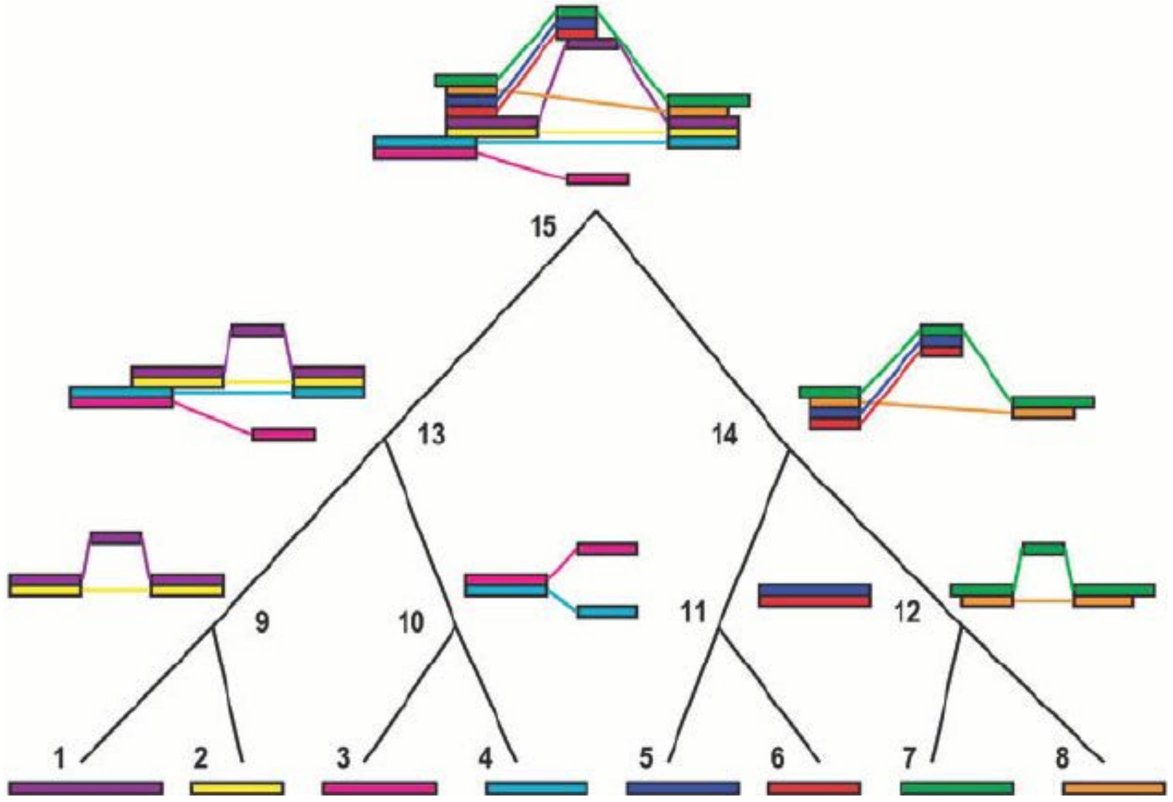
(*i*) Unaligned sequences

(*ii*) Building distance matrix

(*iii*) Guide tree construction

(*iv*) Progressive alignment

(*v*) Aligned sequences



+ efficient enough to work with up to 1000 sequences

– does not provide a global optimal alignment

– errors in the first steps (e.g. erroneous gaps) do propagate to the final alignment

Catherine S Grasso, slide courtesy of Asya Mendelevich

Iterative methods

- works similar to progressive algorithms, but allows to realign the sequences in the alignment on each step
- optimizes a global metric

+ less prone to error propagation, provides a more accurate result

+ works fine with pairwise distant sequences

– still heuristic

– not as efficient as progressive algorithms

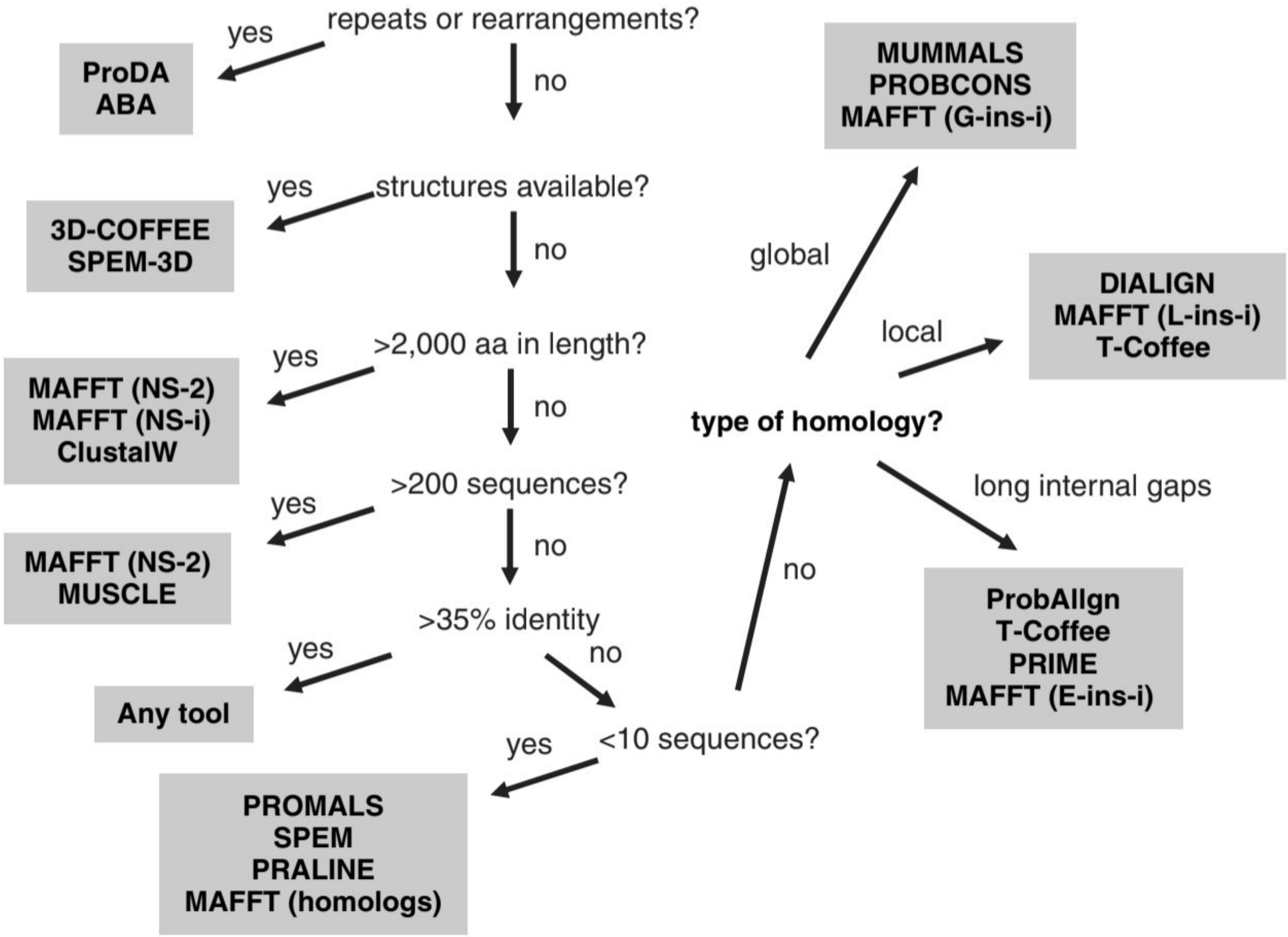
slide courtesy of Asya Mendelevich

Popular aligners

Aligner Algorithm	Type	Input	Comments
MUSCLE	Iterative	DNA, RNA, proteins	Widely used. Allows a lot of options
CLUSTAL Omega	Progressive	DNA, RNA, proteins	O(N log N) guide tree production allows over 100 000 sequences to be aligned. Can reuse existing alignment and append new sequences to them
T-Coffee	Progressive	DNA, RNA, proteins, structures	Wide range of flavors for different situations, e.g. DNA, RNA, proteins. Different modes for fast, accurate, memory-efficient aligning
MAFFT	Iterative	DNA, RNA, proteins	One of the most accurate algorithms for less than 100 sequences. Allows large gaps, making it suitable for rRNA alignments

slide courtesy of Asya Mendelevich

How to select your aligner?



Do and Katoh

Popular aligners

- **ClustalW and ClustalO**
  - documentation, servers and download page: http://www.clustal.org/
  - try: clustalw -INFILE=<fasta> and clustalo --auto --in <fasta> in terminal
- **MUSCLE**
  - documentation and download page: http://www.drive5.com/muscle/
  - server: https://www.ebi.ac.uk/Tools/msa/muscle/
  - try: muscle -in <fasta> in terminal
- **T-Coffee**
  - Coffee family: http://www.tcoffee.org/homepage.html
  - documentation, servers and download page: http://www.tcoffee.org/Projects/tcoffee/
- **MAFFT**
  - documentation, servers and download page: https://mafft.cbrc.jp/alignment/software/

slide courtesy of Asya Mendelevich

How to run aligners?

- Online Tools through **Web Interface**, for small tasks for manual curation:
  - https://www.ebi.ac.uk/Tools/msa/
- **Standalone programs** for larger tasks and manual curation:
  - JalView: https://www.jalview.org/
  - MEGA
- From **bash terminal:** Command Line Interface (CLI), for the large and time-consuming tasks
- From **programming languages**, for full control over input/output:
  - BioPython in Python
  - SciKit-Bio for simple alignments and files parsing in Python
  - msa package for R

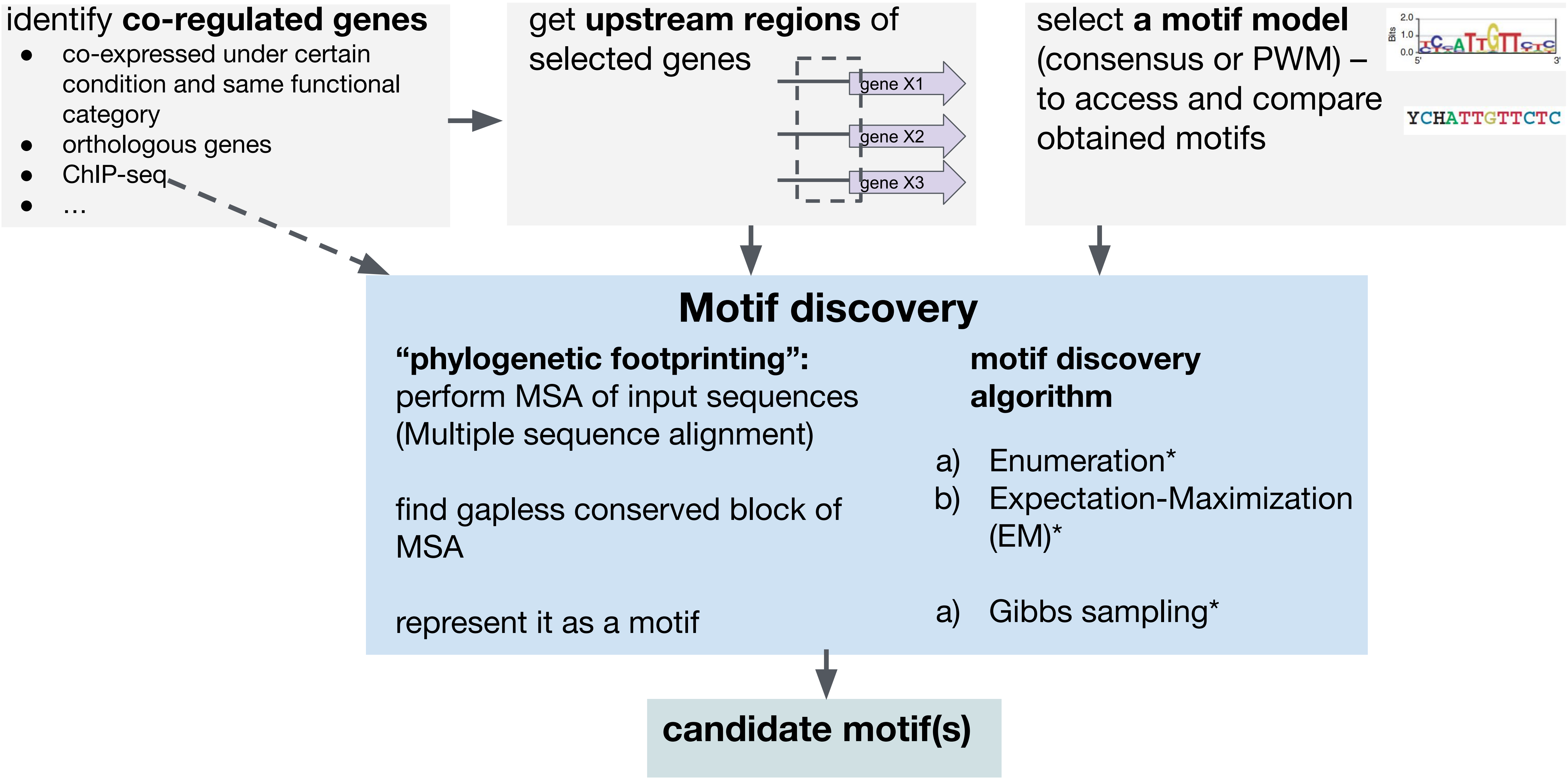
Task 1 (Multiple Alignment)

Instructions:
https://github.com/rybinaanya/2022_Skoltech_Bioinformatics_course_seminar_4

Outline:

Download file with upstream regions of bacterial orthologs `upstreams.fasta`.
Create multiple alignment with T-COFFEE, MUSCLE and CLUSTALW.
Manually select the **most conserved gapless** region and save it into `.fasta` file.

Motif discovery problem: flowchart



identify **co-regulated genes**

- co-expressed under certain condition and same functional category
- orthologous genes
- ChIP-seq
- …

get **upstream regions** of selected genes

select **a motif model** (consensus or PWM) – to access and compare obtained motifs

YCHATTGTTCTC

Motif discovery

“phylogenetic footprinting”: perform MSA of input sequences (Multiple sequence alignment)

find gapless conserved block of MSA

represent it as a motif

motif discovery algorithm

a) Enumeration*
b) Expectation-Maximization (EM)*

a) Gibbs sampling*

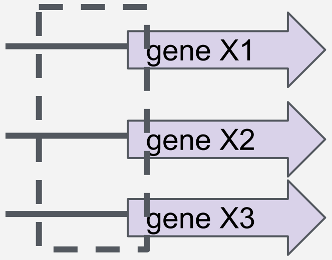
candidate motif(s)

Motif discovery problem: flowchart

identify **co-regulated genes**

- co-expressed under certain condition and same functional category
- orthologeous genes
- ChIP-seq
- …

get **upstream regions** of selected genes



select **a motif model** (consensus or PWM) – to access and compare obtained motifs

2.0
1.0
0.0
Bits
5'
3'

YCHATTGTTCTC

Motif discovery

“phylogenetic footprinting”: perform MSA of input sequences (Multiple sequence alignment)

find gapless conserved block of MSA

represent it as a motif

motif discovery algorithm

a) Enumeration*
b) Expectation-Maximization (EM)*

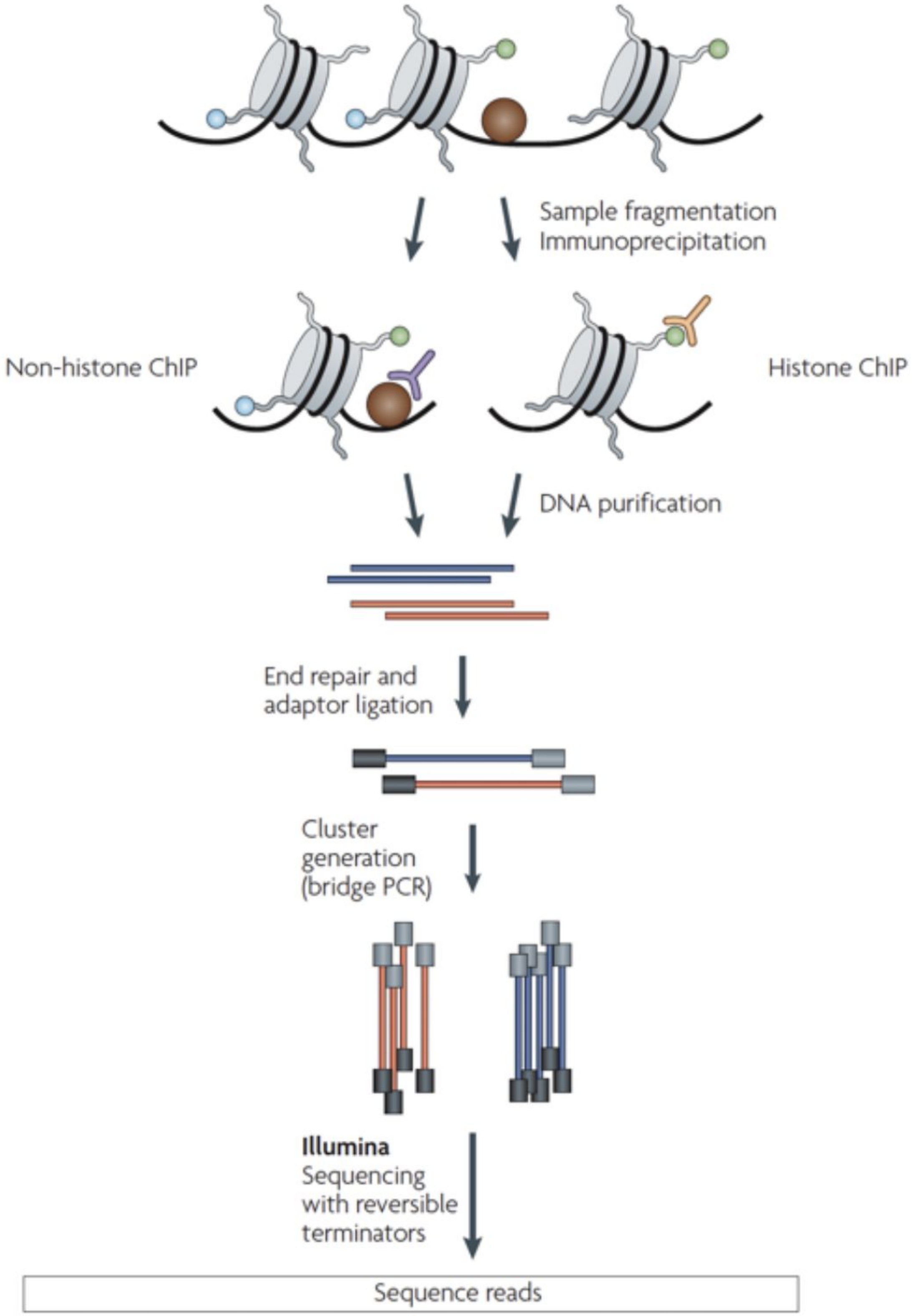
a) Gibbs sampling*

candidate motif(s)

Let’s discuss briefly how we may prepare a set of input sequences using ChIP-seq

ChIP-seq basics: sample preparation for NGS

Chromatin-immunoprecipitation followed by sequencing:



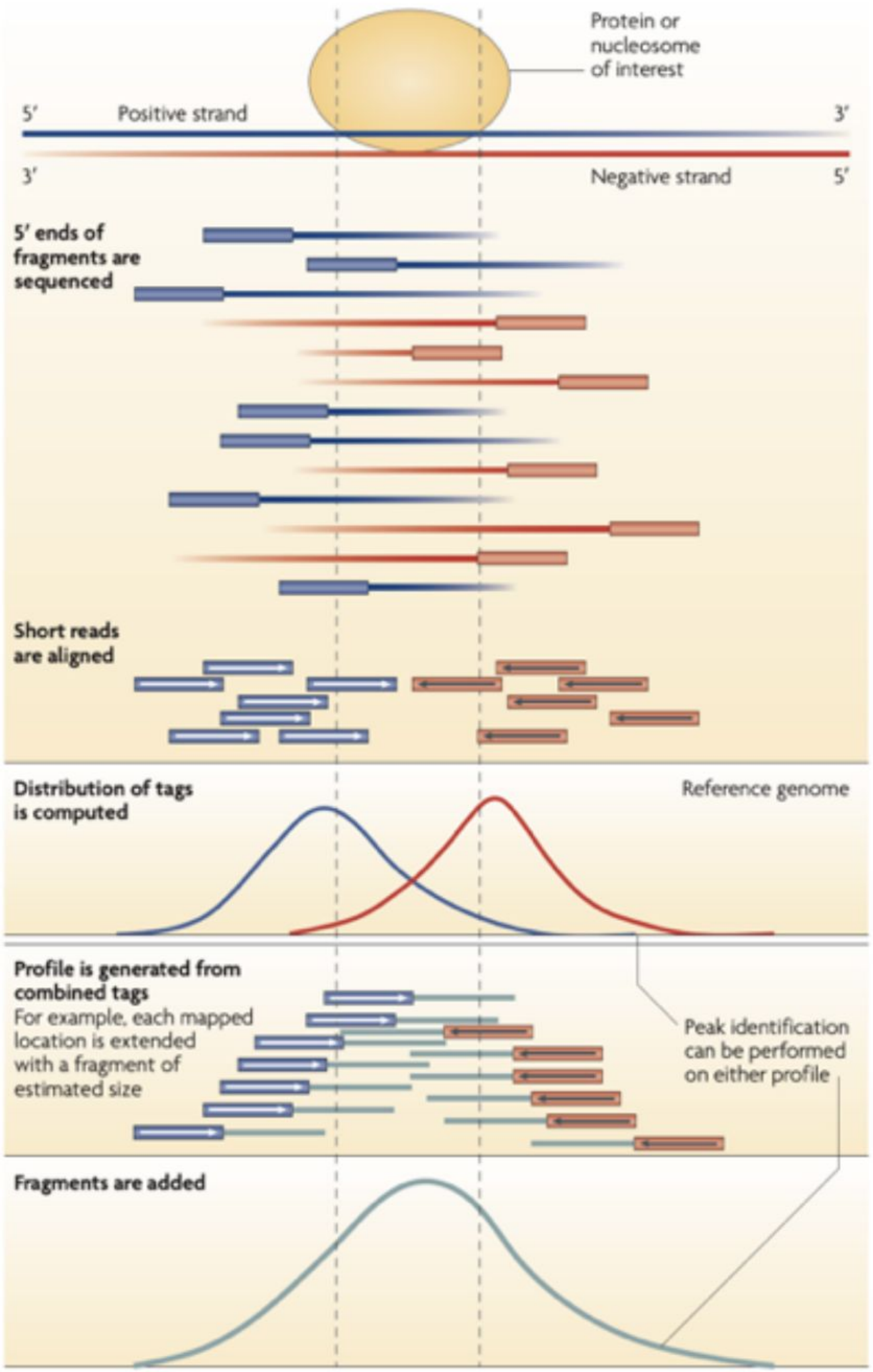
https://www.nature.com/articles/nrg2641

ChIP-seq basics: NGS data analysis

Binding events:

Read alignments:

Peak calling:


Protein or nucleosome of interest
5′
Positive strand
3′
3′
Negative strand
5′
5′ ends of fragments are sequenced
Short reads are aligned
Distribution of tags is computed
Reference genome
Profile is generated from combined tags
For example, each mapped location is extended with a fragment of estimated size
Peak identification can be performed on either profile
Fragments are added

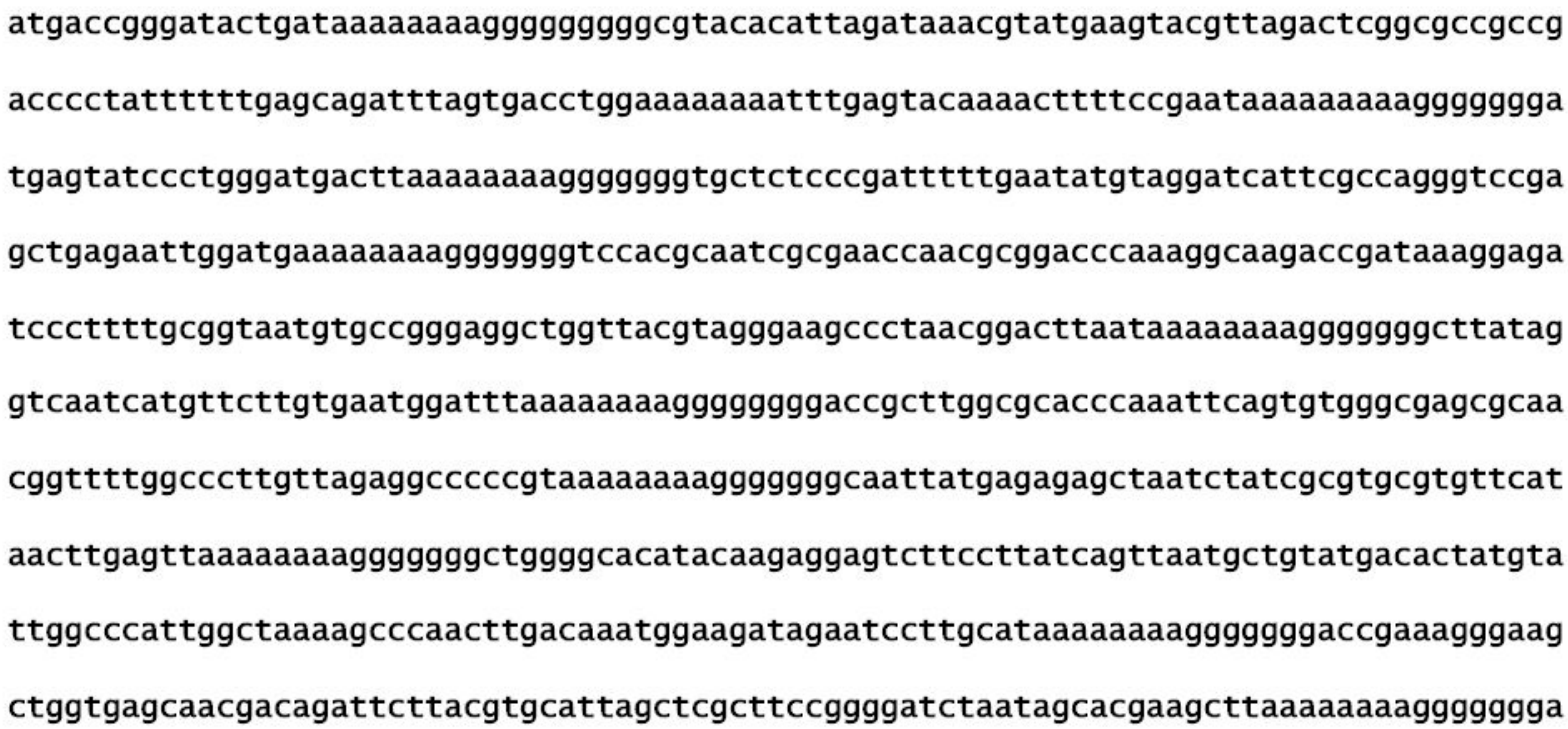
Enroll to the “Omics Data Analysis” course at Skoltech (Term 3-4)

Motif search problem

Given a set of sequences find the motif (number of motifs, the width of each motif and its location in input sequences)

Motif search: training attention

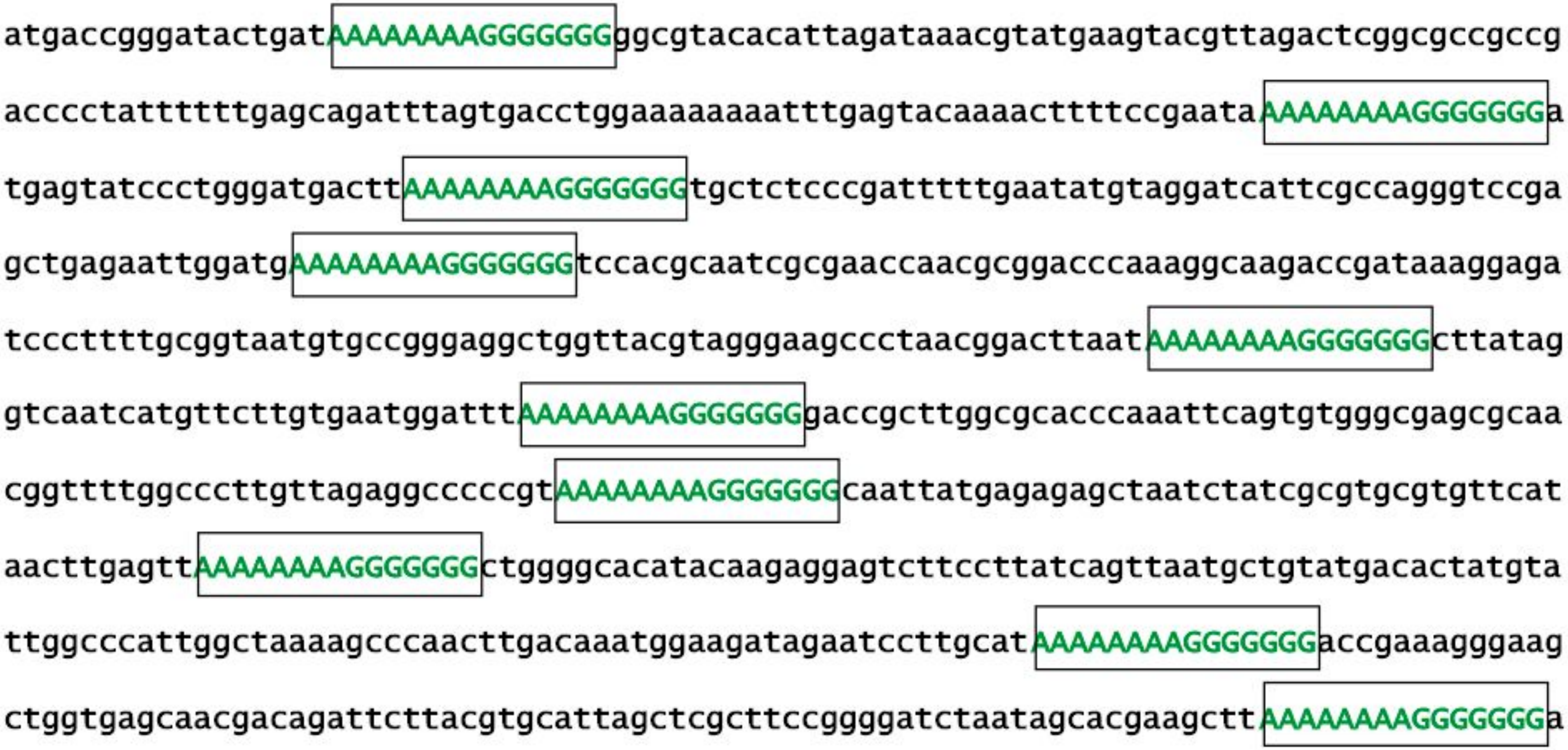
- Try to predict what is the regulatory motif in the following set of sequences:



```
atgaccgggatactgataaaaaaaagggggggggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg

acccctattttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaacttttccgaataaaaaaaaaggggggga

tgagtatccctgggatgacttaaaaaaaagggggggtgctctcccgatttttgaatatgtaggatcattcgccagggtccga

gctgagaattggatgaaaaaaaagggggggtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga

tcccttttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaataaaaaaaagggggggcttatag

gtcaatcatgttcttgtgaatggatttaaaaaaaaggggggggaccgcttggcgcacccaaattcagtgtgggcgagcgcaa

cggttttggcccttgttagaggcccccgtaaaaaaaagggggggcaattatgagagagctaatctatcgcgtgcgtgttcat

aacttgagttaaaaaaaagggggggctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta

ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcataaaaaaaagggggggaccgaaagggaag

ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttaaaaaaaaggggggga
```

Motif search: training attention

- Seems to be easy:



```
atgaccgggatactgatAAAAAAAAGGGGGGGggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg
acccctattttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaacttttccgaataAAAAAAAAGGGGGGGa
tgagtatccctgggatgacttAAAAAAAAGGGGGGGtgctctcccgattttttgaatatgtaggatcattcgccagggtccga
gctgagaattggatgAAAAAAAAGGGGGGGtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga
tcccttttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaatAAAAAAAAGGGGGGGcttatag
gtcaatcatgttcttgtgaatggatttAAAAAAAAGGGGGGGgaccgcttggcgcacccaaattcagtgtgggcgagcgcaa
cggttttggcccttgttagaggcccccgtAAAAAAAAGGGGGGGcaattatgagagagctaatctatcgcgtgcgtgttcat
aacttgagttAAAAAAAAGGGGGGGctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta
ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatAAAAAAAAGGGGGGGaccgaaagggaag
ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttAAAAAAAAGGGGGGGa
```

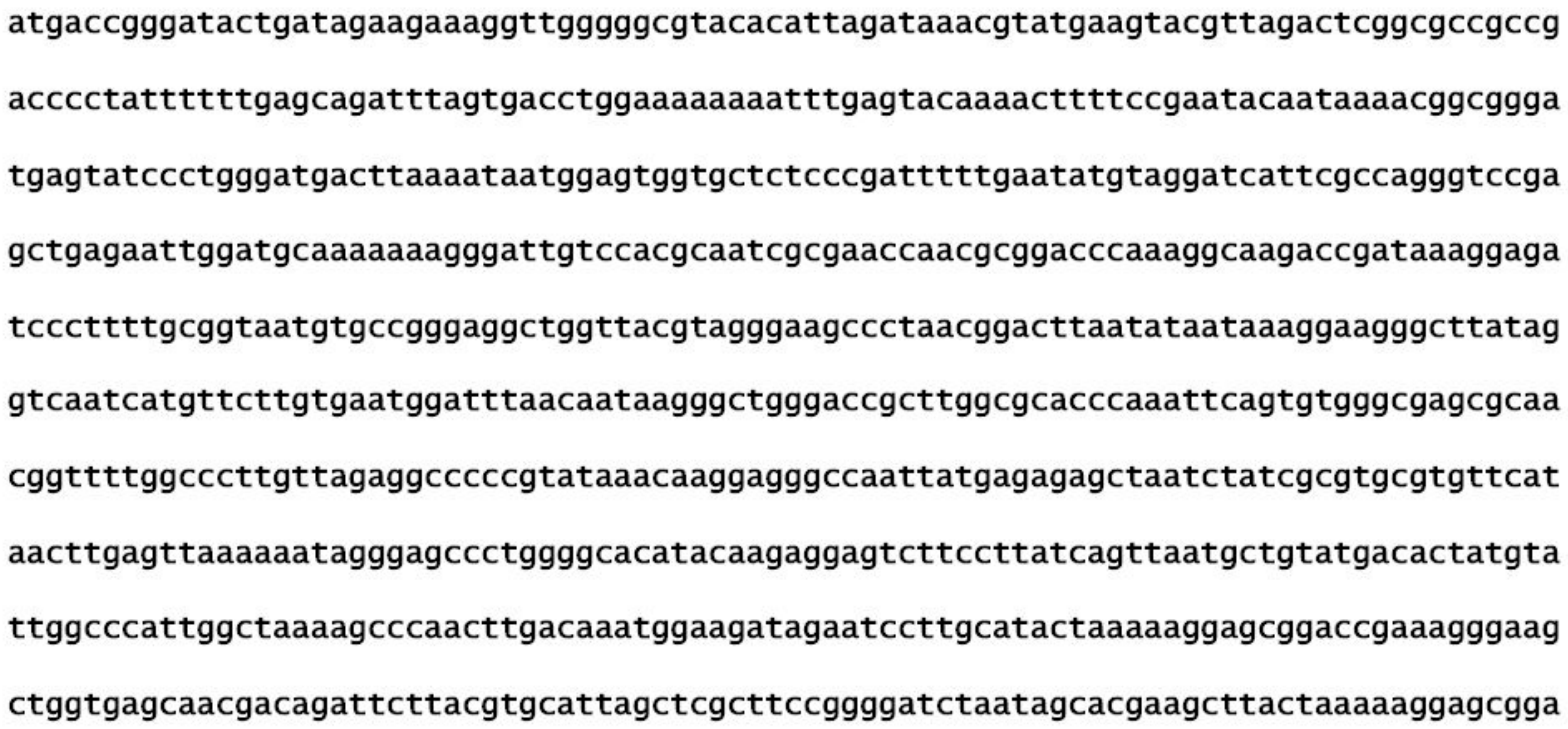
Motif search: training attention

- Let's introduce some substitutions:

```
atgaccgggatactgatAgAAgAAAGGttGGGggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg
acccctattttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaacttttccgaatacAAtAAAAcGGcGGGa
tgagtatccctgggatgacttAAAAtAAtGGaGtGGtgctctcccgatttttgaatatgtaggatcattcgccagggtccga
gctgagaattggatgcAAAAAAAGGGattGtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga
tcccttttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaatAtAAtAAAGGaaGGGcttatag
gtcaatcatgttcttgtgaatggatttAAcAAtAAGGGctGGgaccgcttggcgcacccaaattcagtgtgggcgagcgcaa
cggttttggcccttgttagaggcccccgtAtAAAcAAGGaGGGccaattatgagagagctaatctatcgcgtgcgtgttcat
aacttgagttAAAAAAtAGGGaGccctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta
ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatActAAAAAGGaGcGGaccgaaagggaag
ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttActAAAAAGGaGcGGa
```

Motif search: training attention

- Is everything easy if you know the answer?



```
atgaccgggatactgatagaagaaaggttgggggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg

acccctatttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaacttttccgaatacaataaaacggcggga

tgagtatccctgggatgacttaaaataatggagtggtgctctcccgatttttgaatatgtaggatcattcgccagggtccga

gctgagaattggatgcaaaaaaagggattgtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga

tcccttttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaatataataaaggaagggcttatag

gtcaatcatgttcttgtgaatggatttaacaataagggctgggaccgcttggcgcacccaaattcagtgtgggcgagcgcaa

cggttttggcccttgttagaggcccccgtataaacaaggagggccaattatgagagagctaatctatcgcgtgcgtgttcat

aacttgagttaaaaaatagggagccctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta

ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatactaaaaaggagcggaccgaaagggaag

ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttactaaaaaggagcgga
```

Motif search problem

Given a set of sequences find the motif (number of motifs, the width of each motif and its location in input sequences)

Challenge:

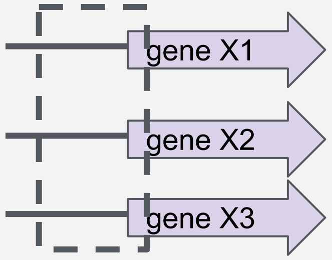
- input sequences could be long (up to thousands and millions)
- motifs are short and could be only slightly similar (due to substitutions)
- we need to distinguish a motif ("signal") from genomic noise (uninformative background DNA)

Motif discovery problem: flowchart

identify **co-regulated genes**

- co-expressed under certain condition and same functional category
- orthologous genes
- ChIP-seq
- …

get **upstream regions** of selected genes



select **a motif model** (consensus or PWM) – to access and compare obtained motifs

2.0
1.0
0.0
Bits
5'
3'

YCHATTGTTCTC

Motif discovery

“phylogenetic footprinting”: perform MSA of input sequences (Multiple sequence alignment)

find gapless conserved block of MSA

represent it as a motif

motif discovery algorithm

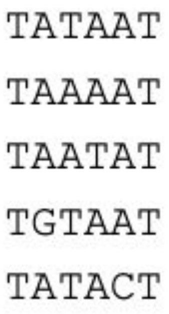
a) Enumeration*
b) Expectation-Maximization (EM)*

a) Gibbs sampling*

candidate motif(s)

To compare, assess, rank motifs, we need a scoring metric and model (way of representation) for motifs

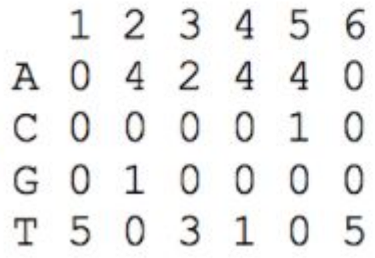
Motif representation (motif models)



```
TATAAT
TAAAAT
TAATAT
TGTAAT
TATACT
```

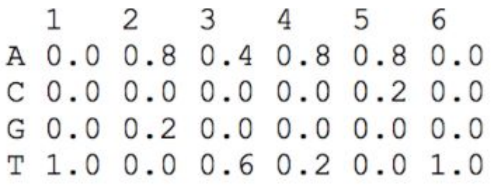
– set of candidate motifs

- consensus sequence – `T[AG][AT][AT][AC]T`
- position frequency matrix (PFM), or position count matrix (PCM) –



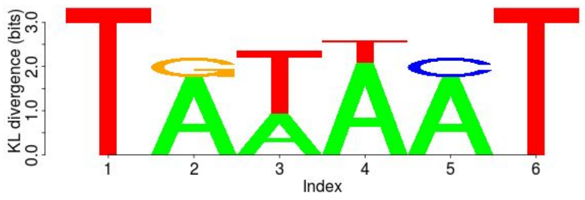
	1	2	3	4	5	6
A	0	4	2	4	4	0
C	0	0	0	0	1	0
G	0	1	0	0	0	0
T	5	0	3	1	0	5

- position probability matrix (PPM) –



	1	2	3	4	5	6
A	0.0	0.8	0.4	0.8	0.8	0.0
C	0.0	0.0	0.0	0.0	0.2	0.0
G	0.0	0.2	0.0	0.0	0.0	0.0
T	1.0	0.0	0.6	0.2	0.0	1.0

- position-specific weight matrix (PWM)
- information content matrix, or sequence logo –



Motif representation: consensus sequence

Consensus sequence lists nucleotides that are allowed in given position. Consider following **gapless block of an alignment**:

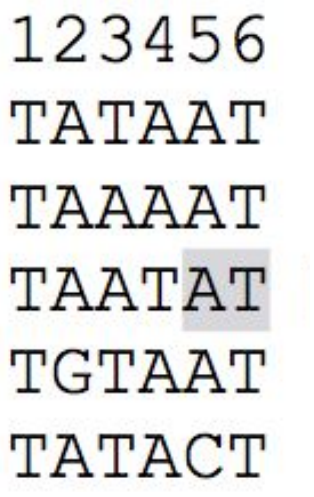
```
TATAAT

TAAAAT

TAATAT

TGTAAT

TATACT
```

Its **consensus**: `T[AG][AT][AT][AC]T`

Problems:

- Doesn't allow to incorporate different preferences for different nucleotides,
- Doesn't allow to account for background nucleotides frequencies.

Motif representation

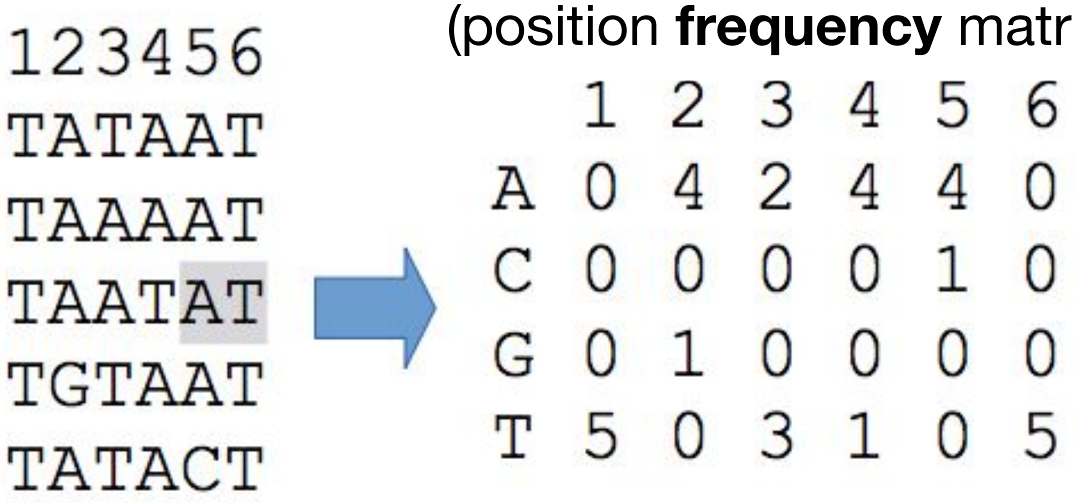


```
123456
TATAAT
TAAAAT
TAATAT
TGTAAT
TATACT
```

Consider a set of candidate motifs obtained from multiple sequence alignment

Motif representation: position **frequency (count)** matrix

position **count** matrix PCM
(position **frequency** matrix PFM)



```
123456
TATAAT
TAAAAT
TAATAT
TGTAAT
TATACT
```

	1	2	3	4	5	6
A	0	4	2	4	4	0
C	0	0	0	0	1	0
G	0	1	0	0	0	0
T	5	0	3	1	0	5

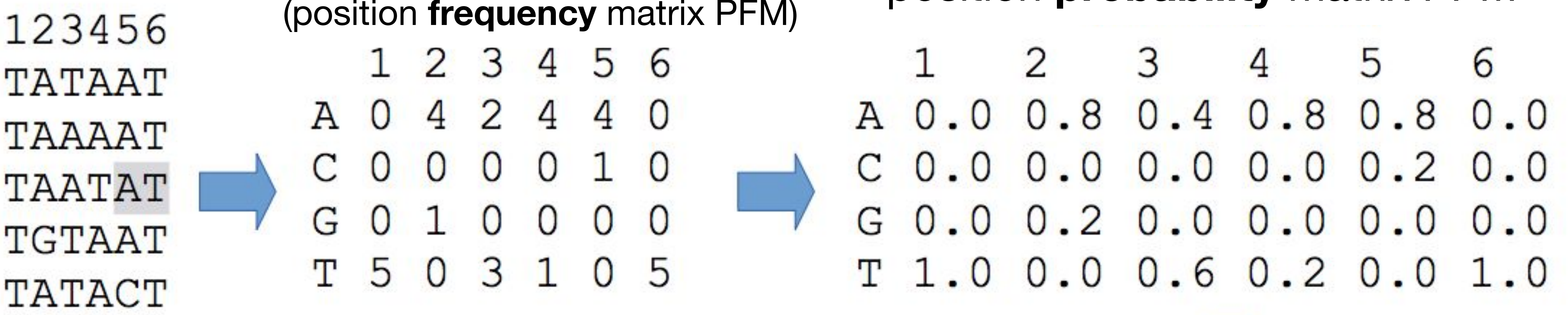
😉 PCM (PFM) counts the **occurrences for each nucleotide in each position**. We have better understanding on different preferences for different nucleotides

😢 PCM (PFM) depends on the number of sequences that were initially aligned

Motif representation: position **probability** matrix

position **count** matrix PCM
(position **frequency** matrix PFM)

position **probability** matrix PPM



```
123456
TATAAT
TAAAAT
TAATAT
TGTAAT
TATACT
```

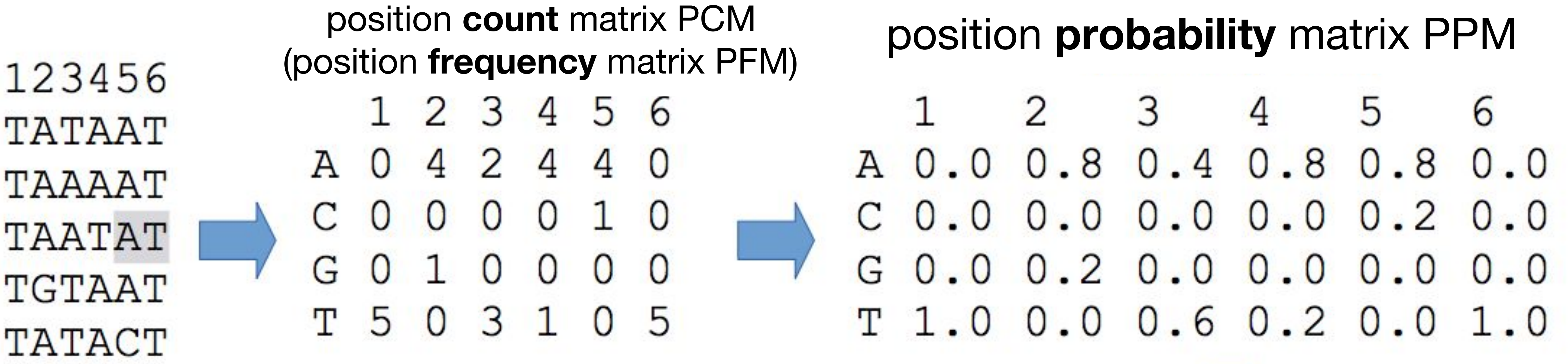
	1	2	3	4	5	6
A	0	4	2	4	4	0
C	0	0	0	0	1	0
G	0	1	0	0	0	0
T	5	0	3	1	0	5

	1	2	3	4	5	6
A	0.0	0.8	0.4	0.8	0.8	0.0
C	0.0	0.0	0.0	0.0	0.2	0.0
G	0.0	0.2	0.0	0.0	0.0	0.0
T	1.0	0.0	0.6	0.2	0.0	1.0

count / (number of input sequences)

😉 PPM **normalizes** the count matrix by the **number of observations**, resulting in an estimate for the probability of a observing each letter at a given position ⇒ Motif representation **no longer depends on the number of sequences aligned**

Motif representation: position **probability** matrix



```
123456
TATAAT
TAAAAT
TAATAT
TGTAAT
TATACT
```

position **count** matrix PCM
(position **frequency** matrix PFM)

	1	2	3	4	5	6
A	0	4	2	4	4	0
C	0	0	0	0	1	0
G	0	1	0	0	0	0
T	5	0	3	1	0	5

position **probability** matrix PPM

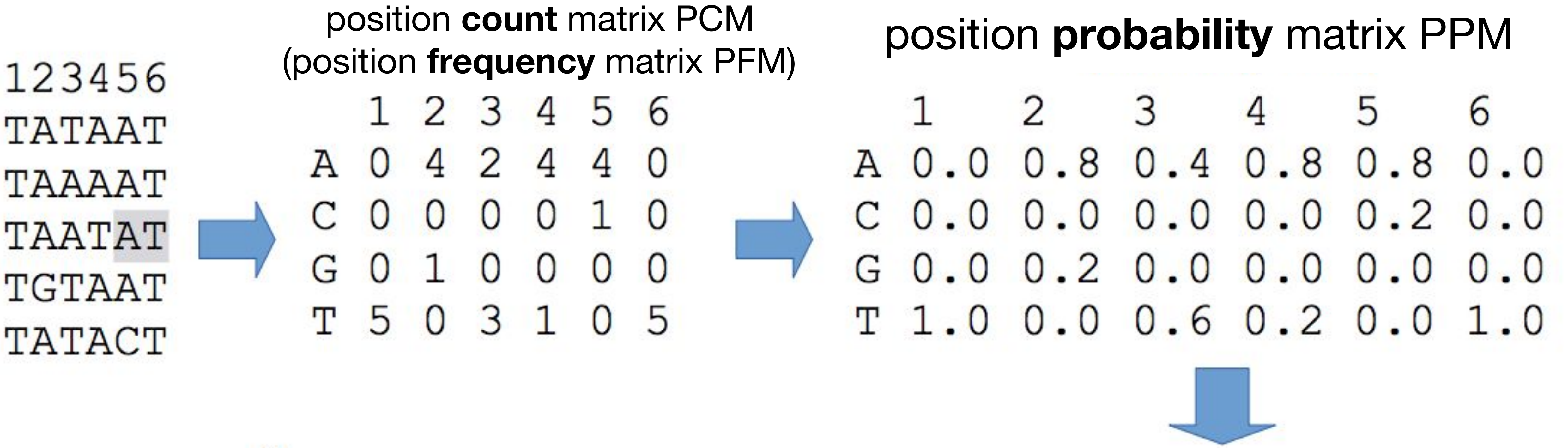
	1	2	3	4	5	6
A	0.0	0.8	0.4	0.8	0.8	0.0
C	0.0	0.0	0.0	0.0	0.2	0.0
G	0.0	0.2	0.0	0.0	0.0	0.0
T	1.0	0.0	0.6	0.2	0.0	1.0

count / (number of input sequences)

😉 PPM **normalizes** the count matrix by the **number of observations**, resulting in an estimate for the probability of a observing each letter at a given position ⇒ Motif representation **no longer depends on the number of sequences aligned**

😢 it does not give us any idea of how “surprising” it would be to observe any given sequence that matches the motif. We need estimate the probability that this observed pattern can be find by chance in the genome. **We need to distinguish informative pattern** (e.g. specific binding, recognized by TF) **from “uninformative” genomic “noise”** (non-specific sites, e.g. not recognized by TF). We need to consider this noise

Motif representation: position **weight** matrix



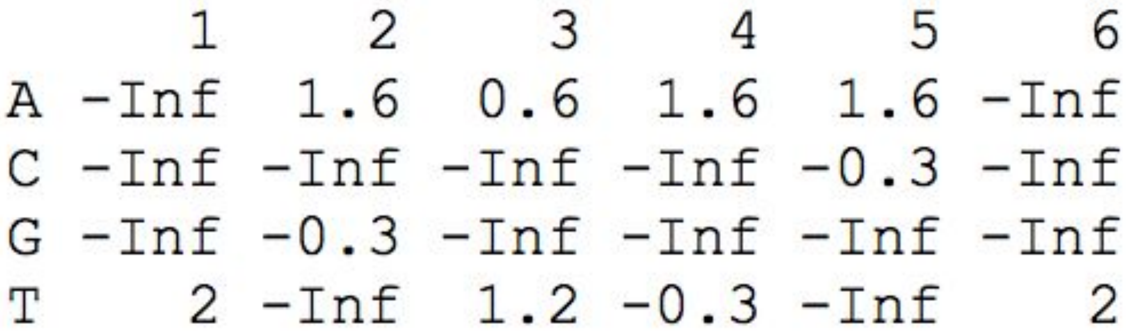
```
123456
TATAAT
TAAAAT
TAATAT
TGTAAT
TATACT
```

position **count** matrix PCM
(position **frequency** matrix PFM)

	1	2	3	4	5	6
A	0	4	2	4	4	0
C	0	0	0	0	1	0
G	0	1	0	0	0	0
T	5	0	3	1	0	5

position **probability** matrix PPM

	1	2	3	4	5	6
A	0.0	0.8	0.4	0.8	0.8	0.0
C	0.0	0.0	0.0	0.0	0.2	0.0
G	0.0	0.2	0.0	0.0	0.0	0.0
T	1.0	0.0	0.6	0.2	0.0	1.0



	1	2	3	4	5	6
A	-Inf	1.6	0.6	1.6	1.6	-Inf
C	-Inf	-Inf	-Inf	-Inf	-0.3	-Inf
G	-Inf	-0.3	-Inf	-Inf	-Inf	-Inf
T	2	-Inf	1.2	-0.3	-Inf	2

$$M_{p,n}=\log 2(\frac{p_{p,n}}{b_n})$$

$p_{p,n}$ is probability of nucleotide *n* in position *p* (column)

b_n is probability of nucleotide *n* in background

"**Background**" refers here to the base composition at **non-specific** sites (i.e. here, sequences that do not necessarily bind the TF). Background is **uniform** (bn = 1/4) or **genome-wide frequencies**

Motif representation: position **weight** matrix

```
123456
TATAAT
TAAAAT
TAATAT
TGTAAT
TATACT
```

→

position **count** matrix PCM
(position **frequency** matrix PFM)

	1	2	3	4	5	6
A	0	4	2	4	4	0
C	0	0	0	0	1	0
G	0	1	0	0	0	0
T	5	0	3	1	0	5

→

position **probability** matrix PPM

	1	2	3	4	5	6
A	0.0	0.8	0.4	0.8	0.8	0.0
C	0.0	0.0	0.0	0.0	0.2	0.0
G	0.0	0.2	0.0	0.0	0.0	0.0
T	1.0	0.0	0.6	0.2	0.0	1.0

↓

	1	2	3	4	5	6
A	-Inf	1.6	0.6	1.6	1.6	-Inf
C	-Inf	-Inf	-Inf	-Inf	-0.3	-Inf
G	-Inf	-0.3	-Inf	-Inf	-Inf	-Inf
T	2	-Inf	1.2	-0.3	-Inf	2

$$M_{p,n} = \log 2(\frac{p_{p,n}}{b_n})$$

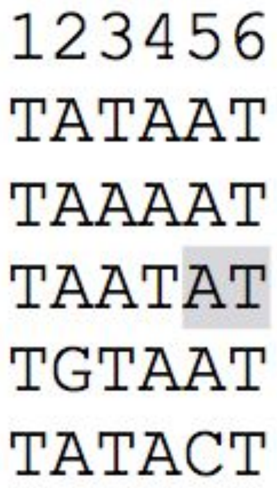
$p_{p,n}$ is probability of nucleotide *n* in position *p* (column)
b_n is probability of nucleotide *n* in background

"**Background**" refers here to the base composition at **non-specific** sites (i.e. here, sequences that do not necessarily bind the TF). Background is **uniform** (bn = 1/4) or **genome-wide frequencies**

😢 But we've got infinity in the matrix!

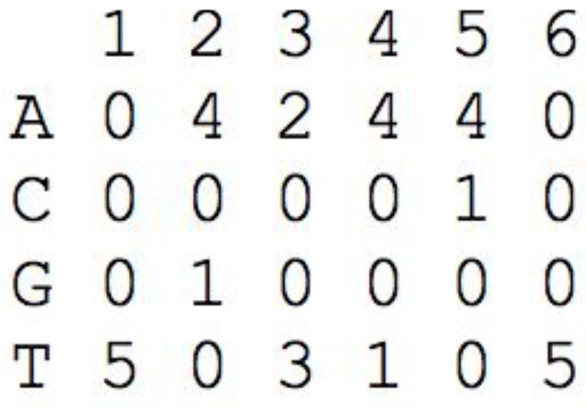
In small datasets, there is always a chance that a possible event does not occur (zeros in count/frequency matrix -> infinity in weight matrix).
To consider rare events and eliminate empirical zero frequencies, we use **pseudocounts.** We will add **pseudocount to each count** in count matrix

Motif representation: position **weight** matrix



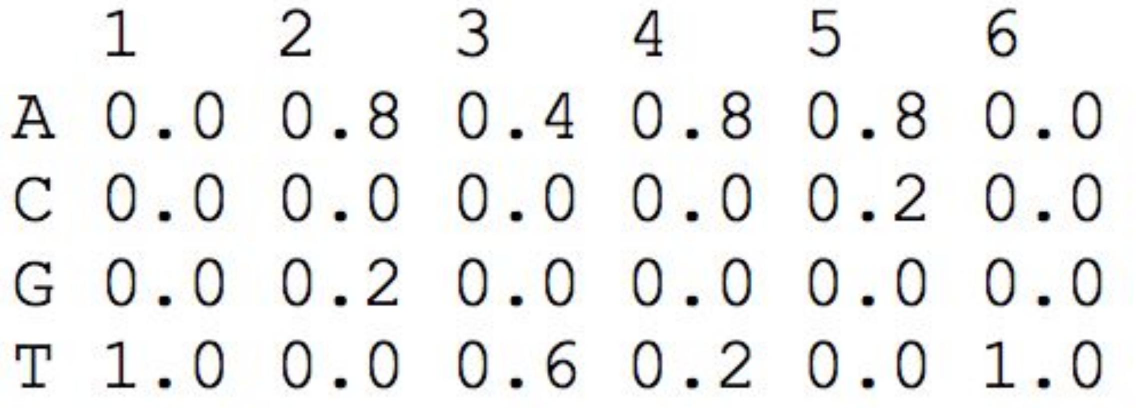
```
123456
TATAAT
TAAAAT
TAATAT
TGTAAT
TATACT
```

position **count** matrix PCM
(position **frequency** matrix PFM)



	1	2	3	4	5	6
A	0	4	2	4	4	0
C	0	0	0	0	1	0
G	0	1	0	0	0	0
T	5	0	3	1	0	5

position **probability** matrix PPM



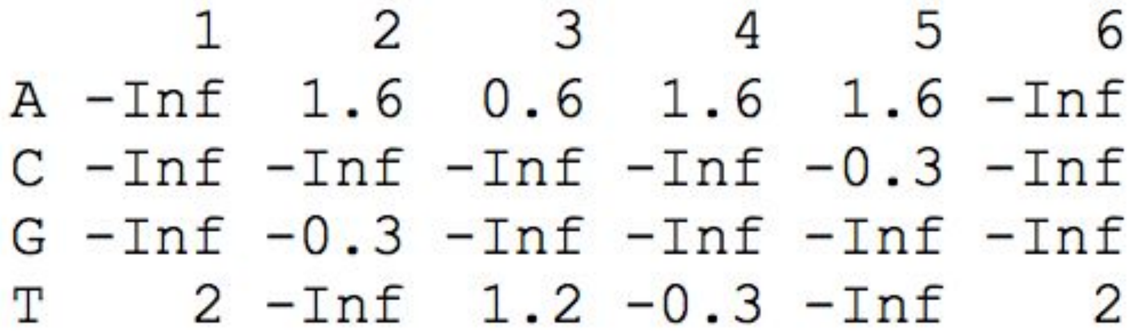
	1	2	3	4	5	6
A	0.0	0.8	0.4	0.8	0.8	0.0
C	0.0	0.0	0.0	0.0	0.2	0.0
G	0.0	0.2	0.0	0.0	0.0	0.0
T	1.0	0.0	0.6	0.2	0.0	1.0

$$M_{p,n} = \log 2\left(\frac{p_{p,n}}{b_n}\right)$$

$p_{p,n}$ is probability of nucleotide *n* in position *p* (column)

b_n is probability of nucleotide *n* in background

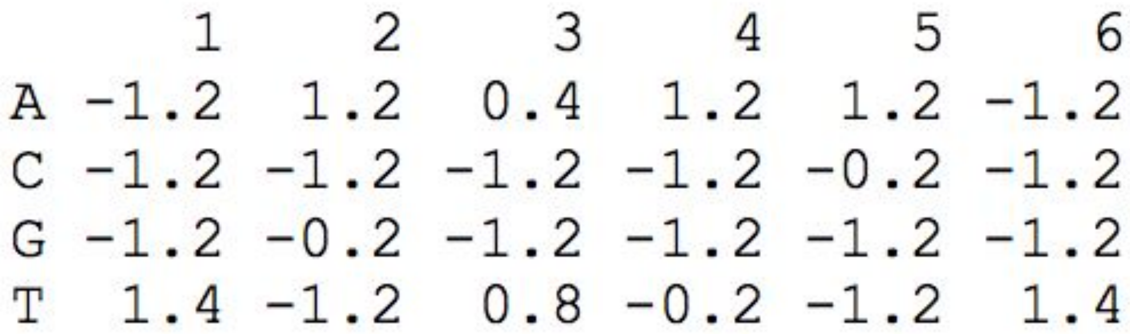
"**Background**" refers here to the base composition at **non-specific** sites (i.e. here, sequences that do not necessarily bind the TF). Background is **uniform** (bn = 1/4) or **genome-wide frequencies**



	1	2	3	4	5	6
A	-Inf	1.6	0.6	1.6	1.6	-Inf
C	-Inf	-Inf	-Inf	-Inf	-0.3	-Inf
G	-Inf	-0.3	-Inf	-Inf	-Inf	-Inf
T	2	-Inf	1.2	-0.3	-Inf	2

position **weight** matrix PWM

Add pseudocounts (for example, 1), to frequency matrix to evade infinity in PWMs. Pseudocounts reflect the fact, that any sequence can be bound by the protein. But some of them are bound with very low probability



	1	2	3	4	5	6
A	-1.2	1.2	0.4	1.2	1.2	-1.2
C	-1.2	-1.2	-1.2	-1.2	-0.2	-1.2
G	-1.2	-0.2	-1.2	-1.2	-1.2	-1.2
T	1.4	-1.2	0.8	-0.2	-1.2	1.4

Motif representation: sequence logo

Relative entropy (Kullback-Leibler distance) of the binding site with respect to the background frequencies:

the frequency of base b at position i

$$I_{seq}(i) = -\sum_{b} f_{b,i} \log_2 \frac{f_{b,i}}{p_b}$$

the background frequency of base b in the genome

Relative entropy measures the degree of disagreement (**dissimilarity**) **between the observed and background base frequencies**, and thus can be used to calculate the significance of the motif itself

https://academic.oup.com/bioinformatics/article/16/1/16/243066

Motif representation: sequence logo

Height of each column is significance of the position (dissimilarity to background).

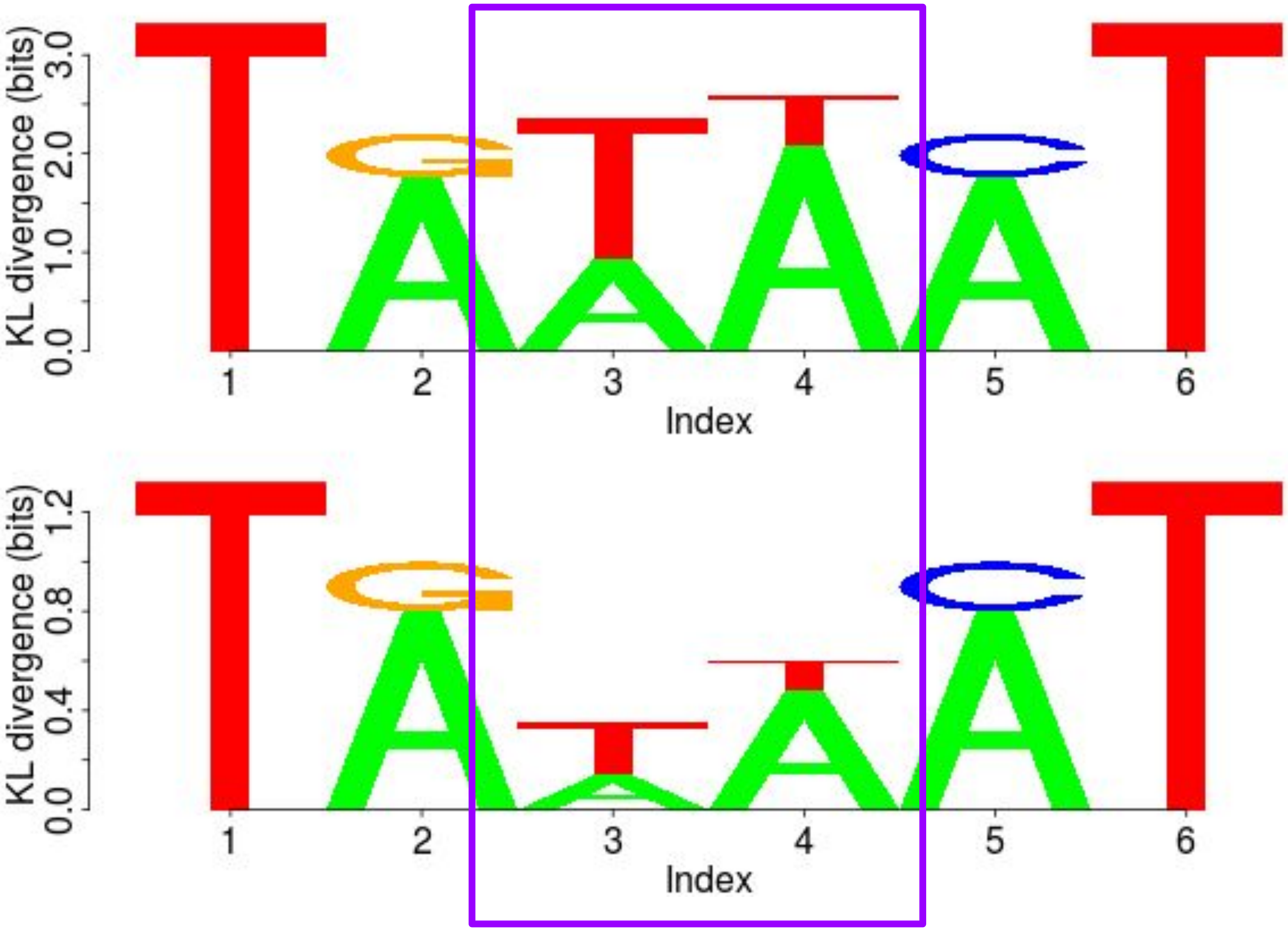
Relative size of the letter is a frequency of the nucleotide.

$$I_{seq}(i) = -\sum_b f_{b,i} \log_2 \frac{f_{b,i}}{p_b}$$

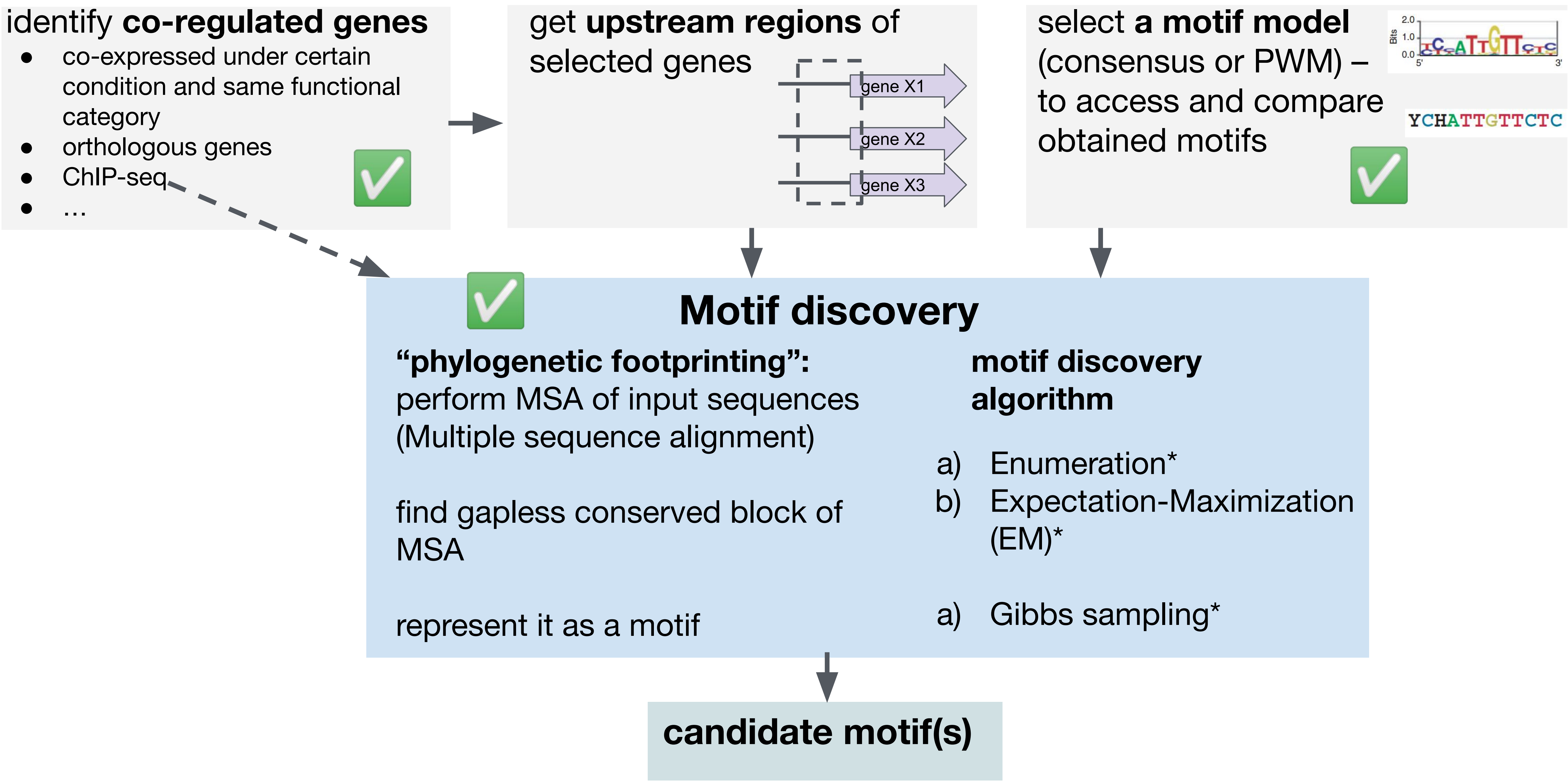
GC-rich background:

AT-rich background:

appearance of T and A is more significant in the GC-rich background than in the AT-rich (=low-GC) background



Motif discovery problem: flowchart



identify **co-regulated genes**

- co-expressed under certain condition and same functional category
- orthologous genes
- ChIP-seq
- …

✅

get **upstream regions** of selected genes

select **a motif model** (consensus or PWM) – to access and compare obtained motifs

YCHATTGTTCTC

✅

Motif discovery

✅

“phylogenetic footprinting”:
perform MSA of input sequences (Multiple sequence alignment)

find gapless conserved block of MSA

represent it as a motif

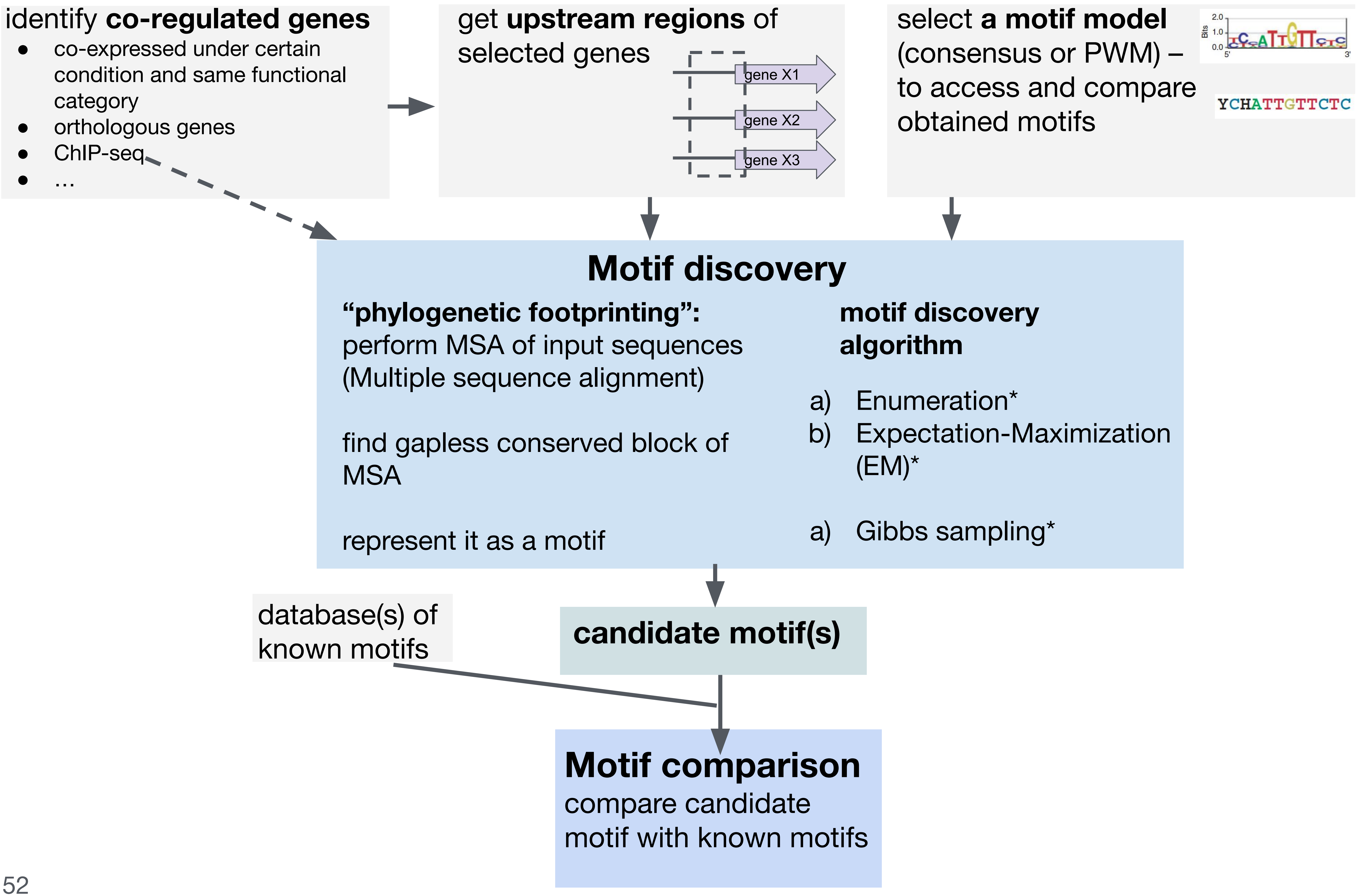
motif discovery algorithm

a) Enumeration*
b) Expectation-Maximization (EM)*

a) Gibbs sampling*

candidate motif(s)

Motif discovery problem: flowchart



identify **co-regulated genes**

- co-expressed under certain condition and same functional category
- orthologous genes
- ChIP-seq
- …

get **upstream regions** of selected genes

select **a motif model** (consensus or PWM) – to access and compare obtained motifs

Motif discovery

“phylogenetic footprinting”: perform MSA of input sequences (Multiple sequence alignment)

find gapless conserved block of MSA

represent it as a motif

motif discovery algorithm

a) Enumeration*
b) Expectation-Maximization (EM)*

a) Gibbs sampling*

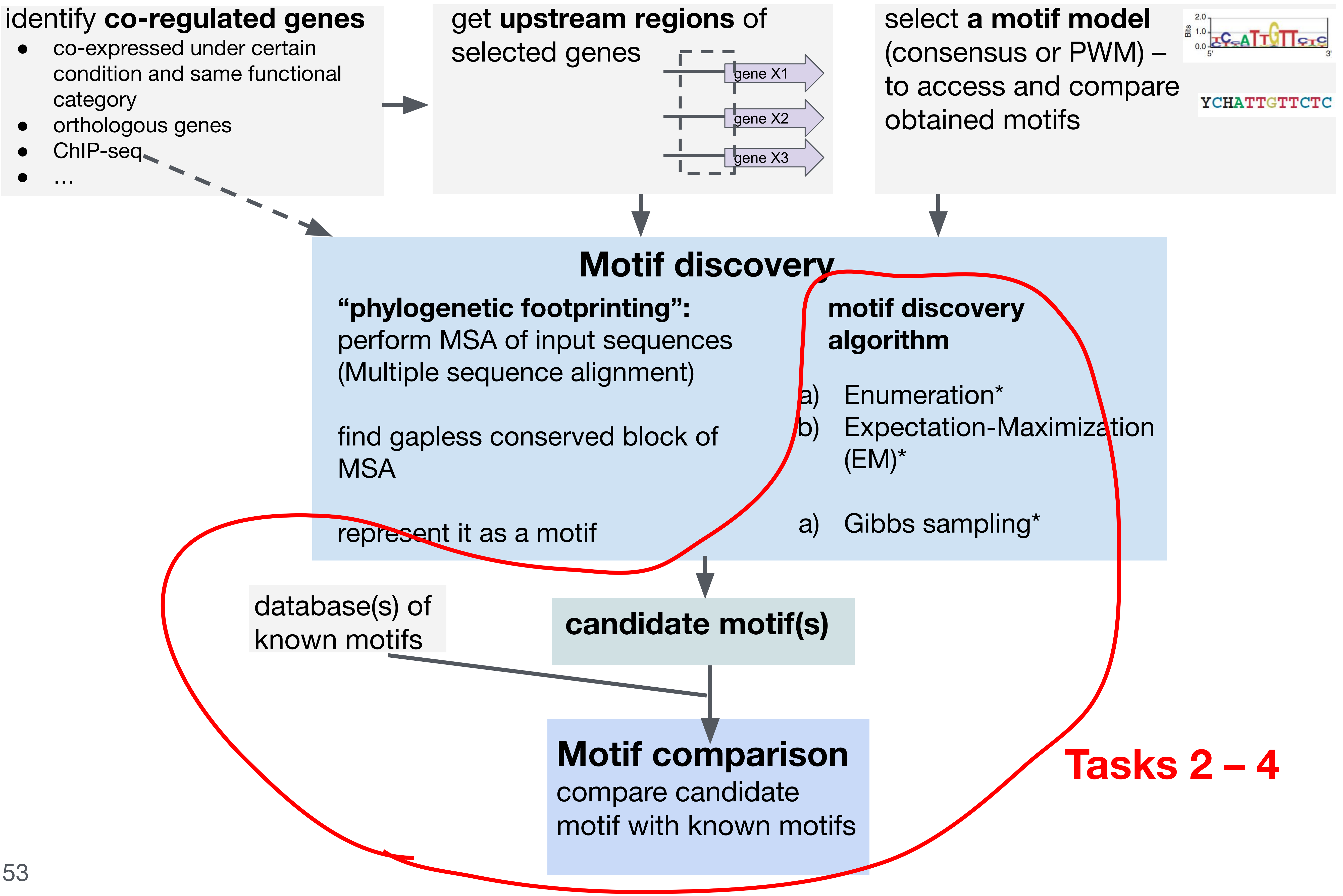
database(s) of known motifs

candidate motif(s)

Motif comparison

compare candidate motif with known motifs

Motif discovery problem: flowchart



identify **co-regulated genes**

- co-expressed under certain condition and same functional category
- orthologous genes
- ChIP-seq
- …

get **upstream regions** of selected genes

select **a motif model** (consensus or PWM) – to access and compare obtained motifs

YCHATTGTTCTC

Motif discovery

“phylogenetic footprinting”: perform MSA of input sequences (Multiple sequence alignment)

find gapless conserved block of MSA

represent it as a motif

motif discovery algorithm

a) Enumeration*
b) Expectation-Maximization (EM)*

a) Gibbs sampling*

database(s) of known motifs

candidate motif(s)

Motif comparison
compare candidate motif with known motifs

Tasks 2 – 4

Some tools for motifs search and manipulation

Web server tools:

- http://rsat.eu/
- http://meme-suite.org/

Console tools:

- https://gimmemotifs.readthedocs.io/en/master/
- http://autosome.ru/

Tools embedded in **programming languages**:

- BioPython motifs

Task 2-4 (Motif search)

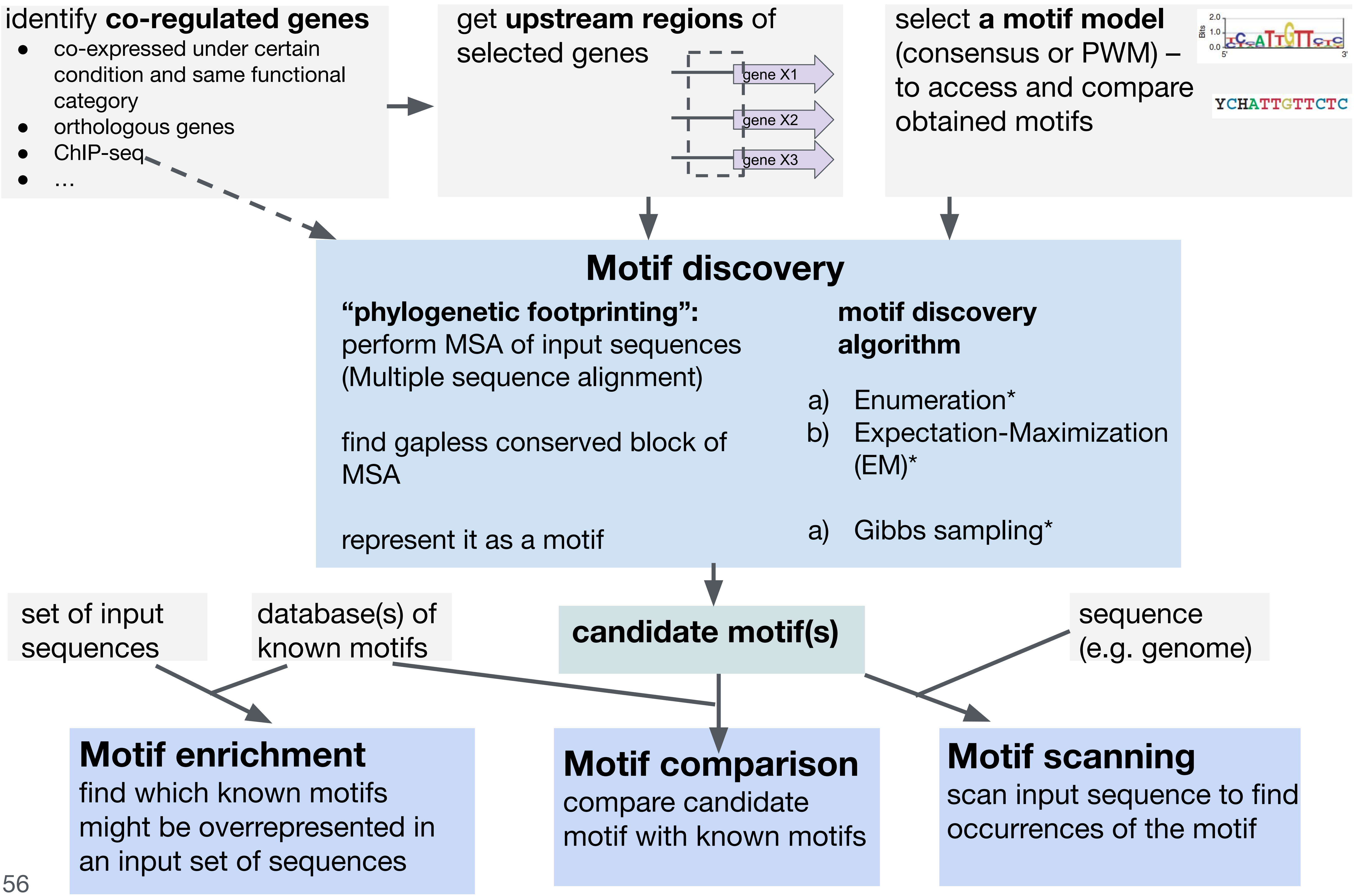
All the **materials** for this seminar are located in Canvas and on GitHub:
https://github.com/rybinaanya/2022_Skoltech_Bioinformatics_course_seminar_4

Go to **instructions**:
https://github.com/rybinaanya/2022_Skoltech_Bioinformatics_course_seminar_4

Outline

- Create counts, frequencies, weights matrices and logo from gapless alignment with RSAT tools: http://embnet.ccg.unam.mx/rsat/ -> Matrix tools.
- Process the same set of sequences `upstreams.fasta` with MEME: http://meme-suite.org/. Set possible length of motif from 5 to 15. Is the result similar to what you found manually?
- Download file with peaks sequences from the given **chicken** ChIP-Seq (`peaks.fasta`).
  Find motifs with MEME-ChIP (http://meme-suite.org/ -> MEME-ChIP).
  What was the protein used for ChIP-Seq?
- Repeat for **your** `peak file` assigned to you in Canvas (see Files for this seminar).

Motif discovery problem: flowchart



identify **co-regulated genes**

- co-expressed under certain condition and same functional category
- orthologous genes
- ChIP-seq
- …

get **upstream regions** of selected genes

select **a motif model** (consensus or PWM) – to access and compare obtained motifs

YCHATTGTTCTC

Motif discovery

"phylogenetic footprinting": perform MSA of input sequences (Multiple sequence alignment)

find gapless conserved block of MSA

represent it as a motif

motif discovery algorithm

a) Enumeration*
b) Expectation-Maximization (EM)*

a) Gibbs sampling*

set of input sequences

database(s) of known motifs

candidate motif(s)

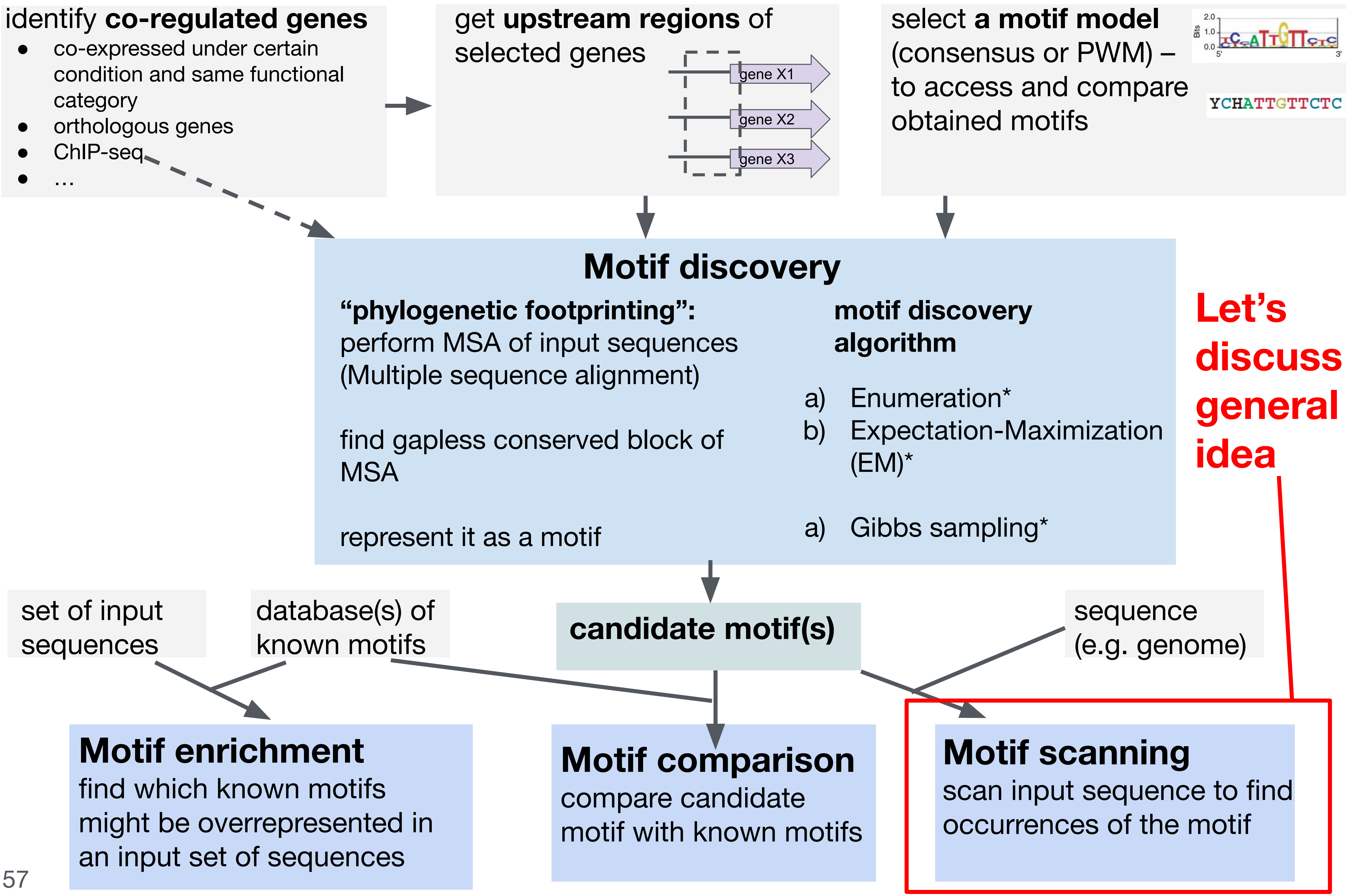
sequence (e.g. genome)

Motif enrichment
find which known motifs might be overrepresented in an input set of sequences

Motif comparison
compare candidate motif with known motifs

Motif scanning
scan input sequence to find occurrences of the motif

Motif discovery problem: flowchart



identify **co-regulated genes**

- co-expressed under certain condition and same functional category
- orthologous genes
- ChIP-seq
- …

get **upstream regions** of selected genes

select **a motif model** (consensus or PWM) – to access and compare obtained motifs

YCHATTGTTCTC

Motif discovery

“phylogenetic footprinting”: perform MSA of input sequences (Multiple sequence alignment)

find gapless conserved block of MSA

represent it as a motif

motif discovery algorithm

a) Enumeration*
b) Expectation-Maximization (EM)*

a) Gibbs sampling*

Let’s discuss general idea

set of input sequences

database(s) of known motifs

candidate motif(s)

sequence (e.g. genome)

Motif enrichment
find which known motifs might be overrepresented in an input set of sequences

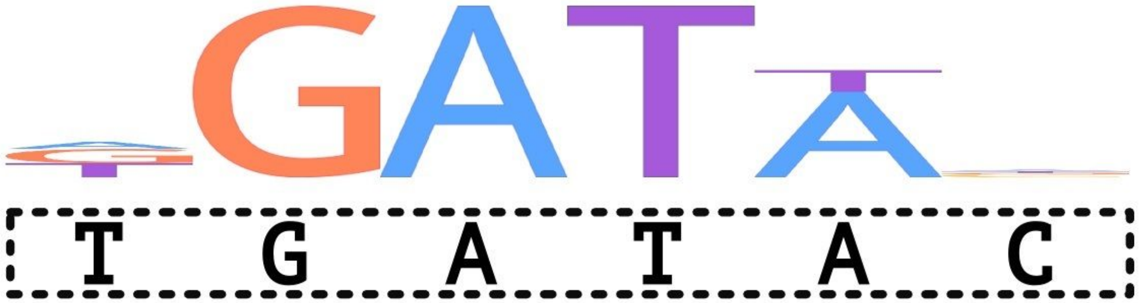
Motif comparison
compare candidate motif with known motifs

Motif scanning
scan input sequence to find occurrences of the motif

Motif scanning: estimate how well input sequence matches the motif

- Let's imagine that we know particular motif and its PWM for some protein. How can we find the binding sites of this protein in the genome?

input data:



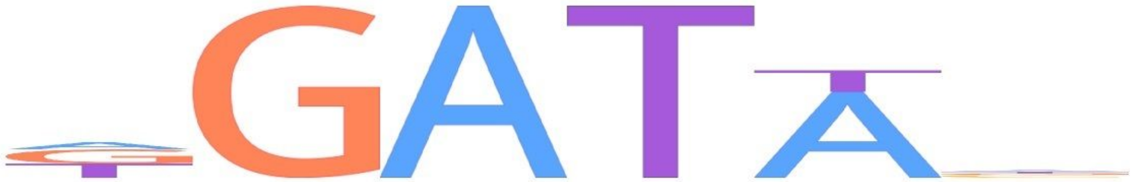
T G A T A C

motif logo

given sequence
(e.g. genome)

Motif scanning: estimate how well input sequence matches the motif

- Let's imagine that we know particular motif and its PWM for some protein. How can we find the binding sites of this protein in the genome?



motif logo

input data:

T G A T A C

given sequence (e.g. genome)

For each position of the PWM, highlight score of the matching letter

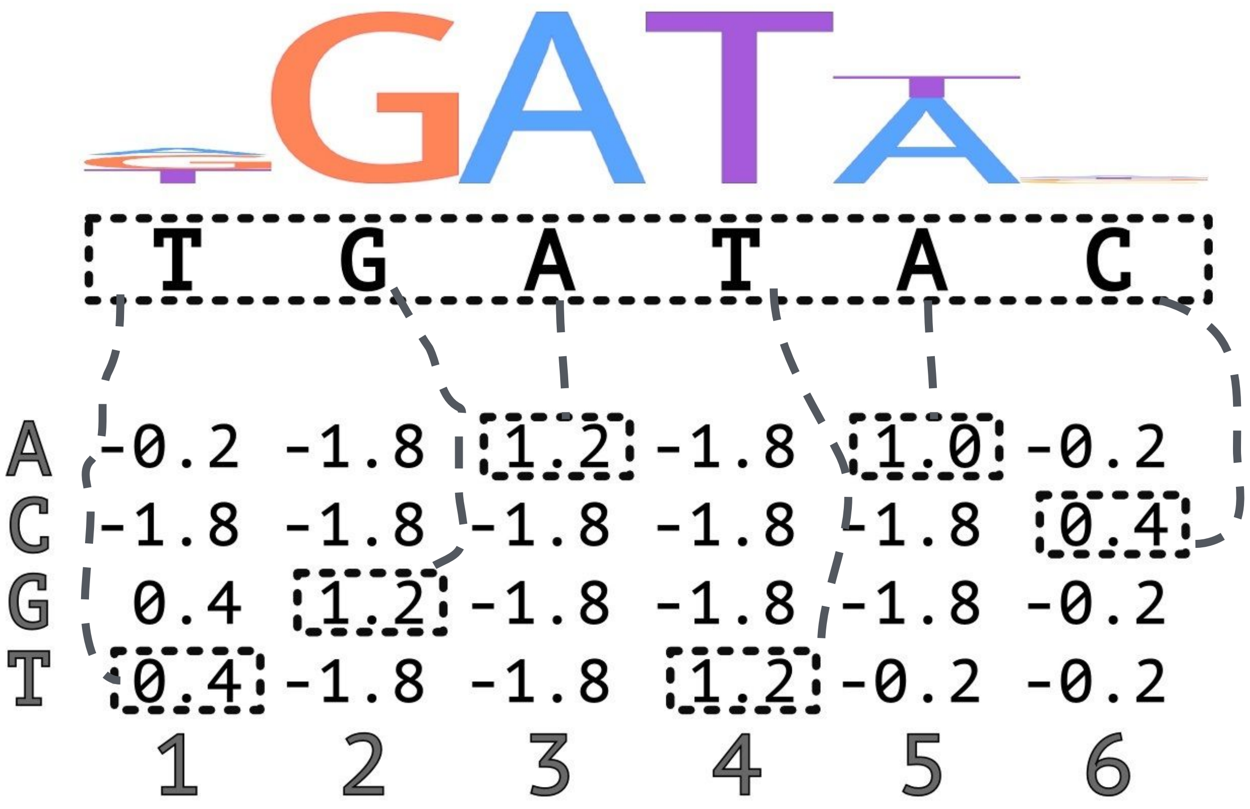
A	-0.2	-1.8	1.2	-1.8	1.0	-0.2
C	-1.8	-1.8	-1.8	-1.8	-1.8	0.4
G	0.4	1.2	-1.8	-1.8	-1.8	-0.2
T	0.4	-1.8	-1.8	1.2	-0.2	-0.2
	1	2	3	4	5	6

motif PWM

Motif scanning: estimate how well input sequence matches the motif

- Let's imagine that we know particular motif and its PWM for some protein. How can we find the binding sites of this protein in the genome?

input data:



motif logo

given sequence (e.g. genome)

T G A T A C

For each position of the PWM, highlight score of the matching letter

	1	2	3	4	5	6
A	-0.2	-1.8	1.2	-1.8	1.0	-0.2
C	-1.8	-1.8	-1.8	-1.8	-1.8	0.4
G	0.4	1.2	-1.8	-1.8	-1.8	-0.2
T	0.4	-1.8	-1.8	1.2	-0.2	-0.2

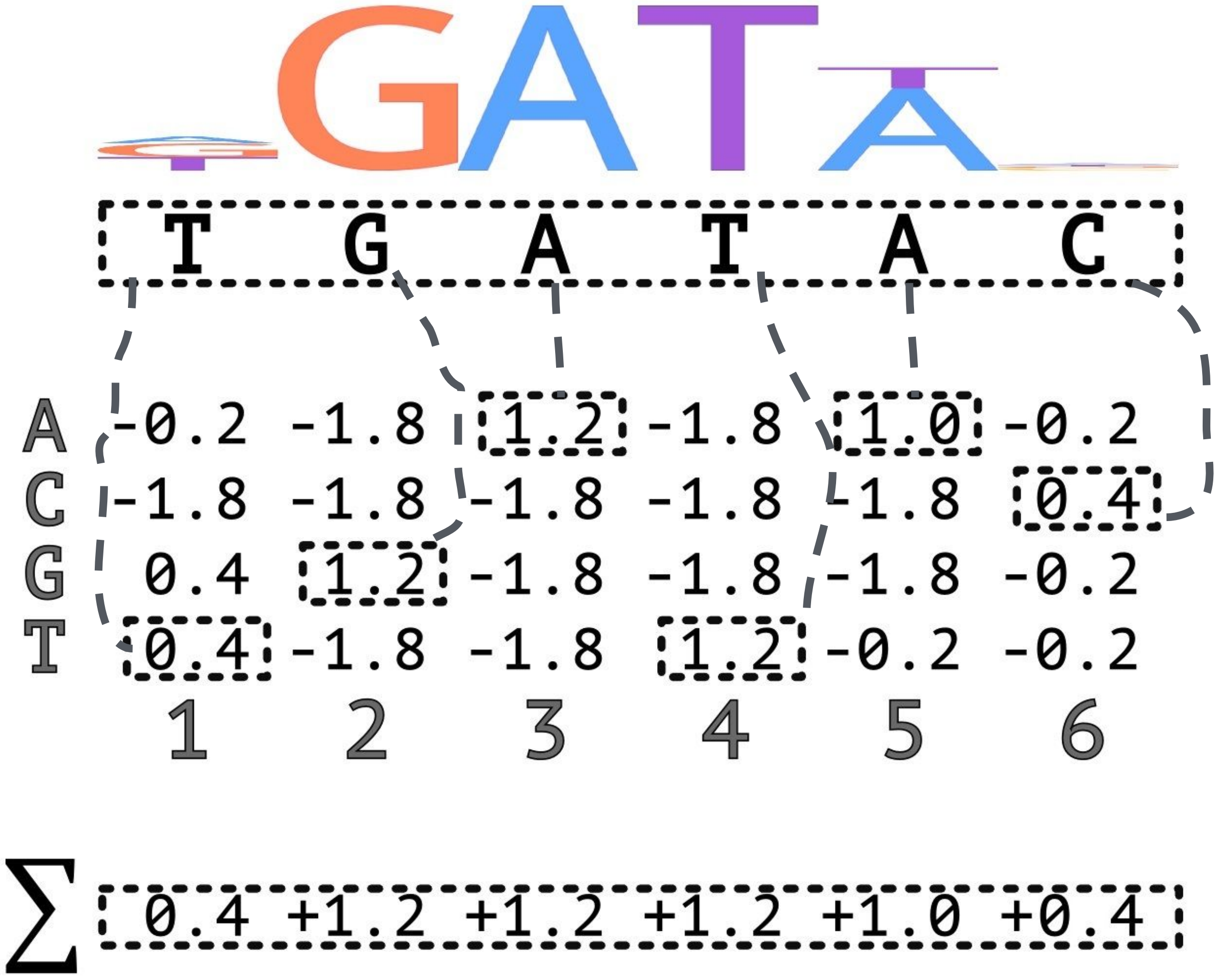
motif PWM

Motif scanning: estimate how well input sequence matches the motif

- Let's imagine that we know particular motif and its PWM for some protein. How can we find the binding sites of this protein in the genome?

input data:

motif logo



given sequence (e.g. genome)

For each position of the PWM, highlight score of the matching letter

motif PWM

	1	2	3	4	5	6
A	-0.2	-1.8	1.2	-1.8	1.0	-0.2
C	-1.8	-1.8	-1.8	-1.8	-1.8	0.4
G	0.4	1.2	-1.8	-1.8	-1.8	-0.2
T	0.4	-1.8	-1.8	1.2	-0.2	-0.2

A sequences’ score for a given motif represents how well the sequence matches the motif

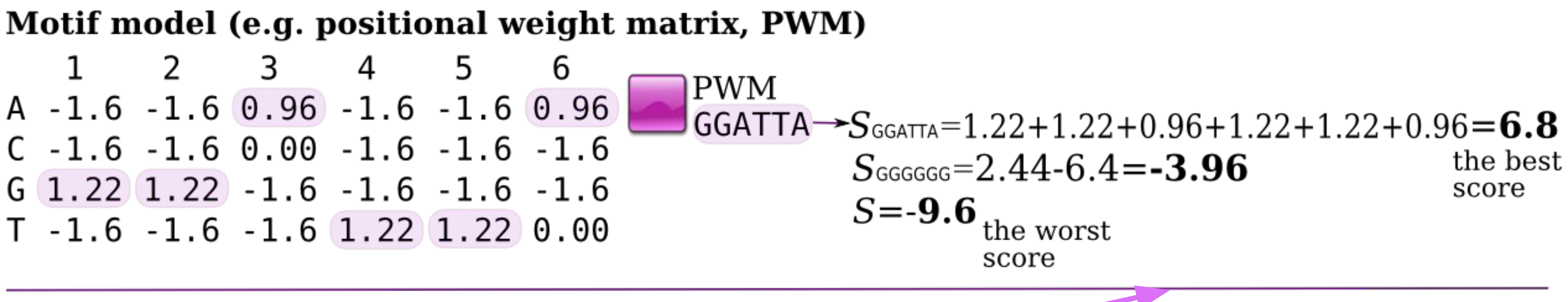
score

$$\sum 0.4 + 1.2 + 1.2 + 1.2 + 1.0 + 0.4$$

$$\textit{Score}(\text{tGATAc}) = 5.4$$

- Additivity assumption: score is larger for longer sequences!

Motif scanning: select the score threshold to mark candidate binding sites



Motif model (e.g. positional weight matrix, PWM)

	1	2	3	4	5	6
A	-1.6	-1.6	0.96	-1.6	-1.6	0.96
C	-1.6	-1.6	0.00	-1.6	-1.6	-1.6
G	1.22	1.22	-1.6	-1.6	-1.6	-1.6
T	-1.6	-1.6	-1.6	1.22	1.22	0.00

PWM

GGATTA → $S_{GGATTA}=1.22+1.22+0.96+1.22+1.22+0.96=\mathbf{6.8}$ the best score

$S_{GGGGGG}=2.44\text{-}6.4=\mathbf{-3.96}$

$S=\mathbf{-9.6}$ the worst score

a) S_min as threshold:

False positive: S_GGGGGG=-3.96 > S_min = -9.6 ⇒ S_GGGGGG has passed but it is not a true motif! not cool

a) S_max as threshold:

S_GGGGGG=-3.96 < S_max = 6.8 ⇒ S_GGGGGG is rejected and it is not true motif, everything is ok

Motif scanning: select the score threshold to mark candidate binding sites

Motif model (e.g. positional weight matrix, PWM)

	1	2	3	4	5	6
A	-1.6	-1.6	0.96	-1.6	-1.6	0.96
C	-1.6	-1.6	0.00	-1.6	-1.6	-1.6
G	1.22	1.22	-1.6	-1.6	-1.6	-1.6
T	-1.6	-1.6	-1.6	1.22	1.22	0.00

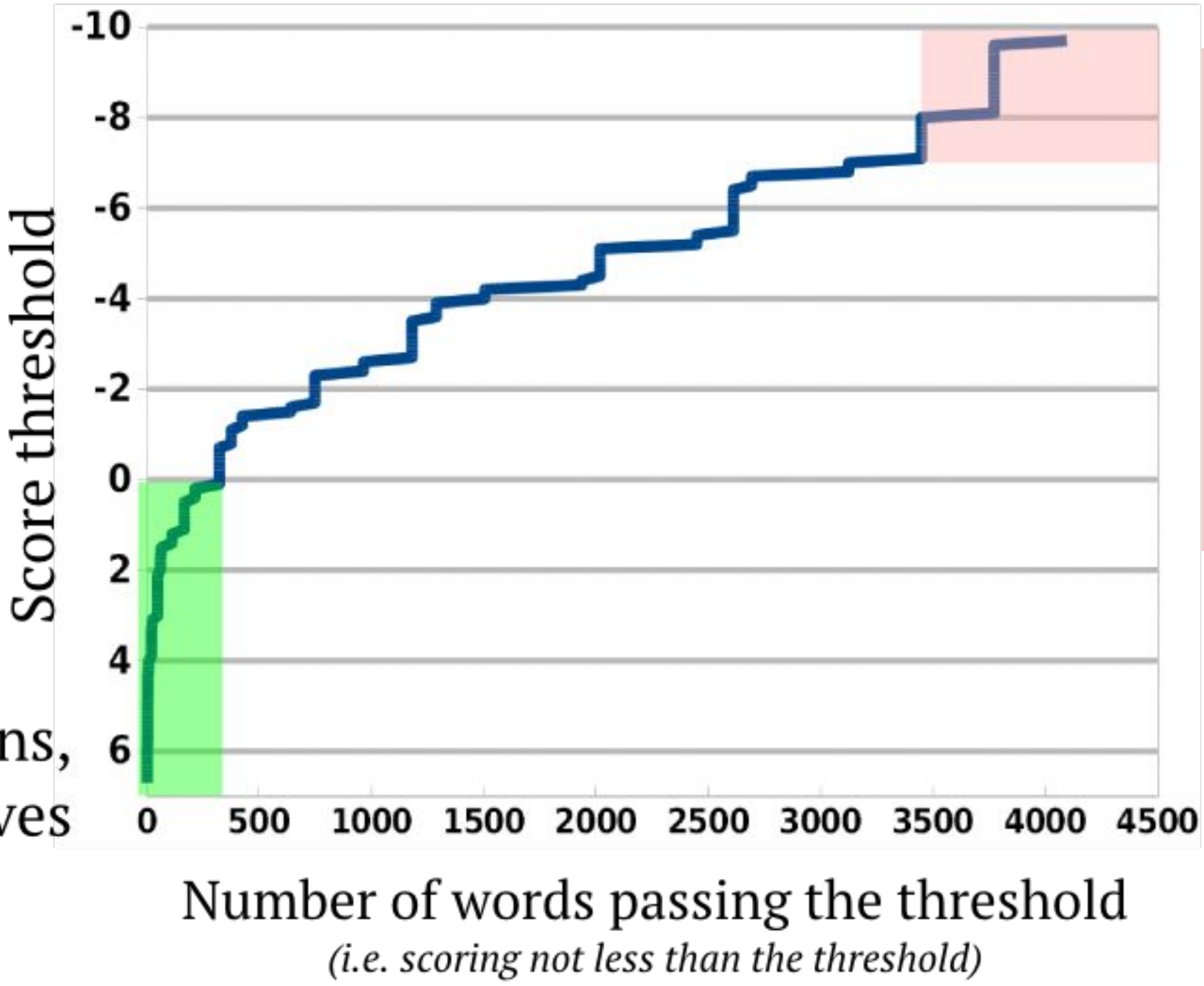
PWM

GGATTA → $S_{GGATTA}=1.22+1.22+0.96+1.22+1.22+0.96=\mathbf{6.8}$ the best score

$S_{GGGGGG}=2.44-6.4=\mathbf{-3.96}$

$S=\mathbf{-9.6}$ the worst score

more predicted TFBS,
more false positive predictions



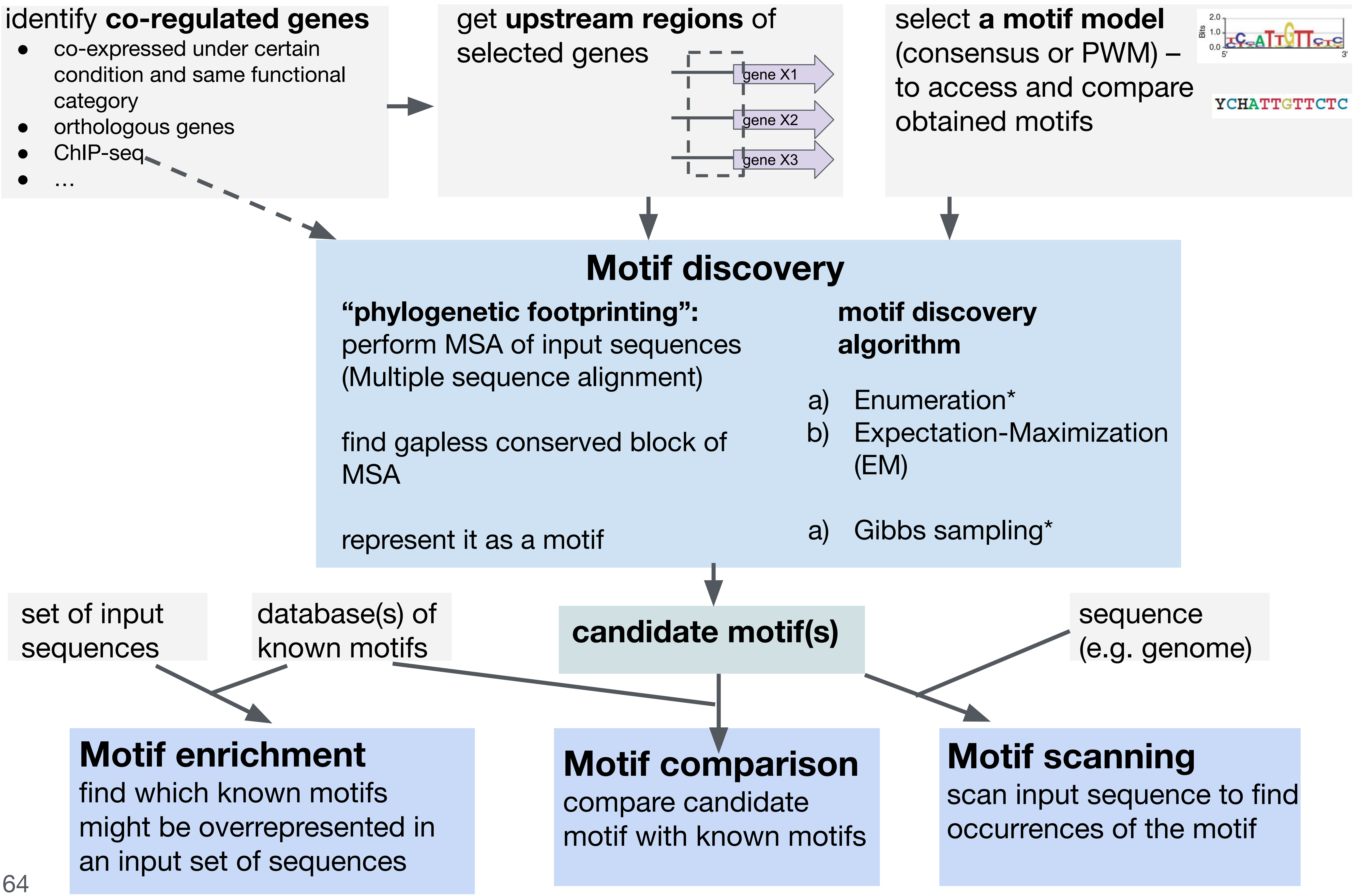
threshold is too strict, too high
a lot of patterns that are real motifs do not pass the threshold and we lose them

threshold is too negative, too low
a lot of patterns pass a threshold and are reported as motifs but they are not real motifs

less TFBS predictions,
less true positives

Score threshold turns a motif model into a binary "yes/no" classifier!

Motif discovery problem: flowchart



identify **co-regulated genes**

- co-expressed under certain condition and same functional category
- orthologous genes
- ChIP-seq
- …

get **upstream regions** of selected genes

select **a motif model** (consensus or PWM) – to access and compare obtained motifs

Motif discovery

“phylogenetic footprinting”: perform MSA of input sequences (Multiple sequence alignment)

find gapless conserved block of MSA

represent it as a motif

motif discovery algorithm

a) Enumeration*
b) Expectation-Maximization (EM)

a) Gibbs sampling*

set of input sequences

database(s) of known motifs

candidate motif(s)

sequence (e.g. genome)

Motif enrichment
find which known motifs might be overrepresented in an input set of sequences

Motif comparison
compare candidate motif with known motifs

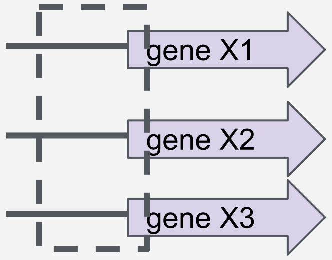
Motif scanning
scan input sequence to find occurrences of the motif

Search for regulatory **DNA** motifs: computational approach

identify **co-regulated genes**

- co-expressed under certain condition and same functional category
- orthologous genes
- ChIP-seq
- …

get **upstream regions** of selected genes



select **a motif model** (consensus or PWM) – to access and compare obtained motifs

2.0
1.0
0.0
Bits
5'
3'

YCHATTGTTCTC

Motif discovery

"phylogenetic footprinting":
perform MSA of input sequences (Multiple sequence alignment)

find gapless conserved block of MSA

represent it as a motif

MSA:
Muscle/T-coffee/ClustalW
Visualization: MView

RSAT RSAT Metazoa

motif discovery algorithm

a) Enumeration*
b) Expectation-Maximization (EM) MEME Multiple Em for Motif Elicitation

a) Gibbs sampling*

candidate motif(s)

set of input sequences

database(s) of known motifs

sequence (e.g. genome)

Motif enrichment

find which known motifs might be overrepresented in an input set of sequences

Tomtom Motif Comparison Tool

Motif comparison

compare candidate motif with known motifs

Motif scanning

scan input sequence to find occurrences of the motif

Homework

Assignment is in Canvas (quizz format)

Files for assignment – Canvas and github

Instructions:
https://github.com/rybinaanya/2022_Skoltech_Bioinformatics_course_seminar_4

Deadline: 12:00 (midday), 23 November Wed

If you have questions, please e-mail me **anna.rybina@skoltech.ru**

Useful links for future learning

Multiple Sequence Alignment Methods Edited by David J. Russell.
https://doi.org/10.1007/978-1-62703-646-7

Multiple Sequence Alignment Edited by Kazutaka Katoh.
https://link.springer.com/book/10.1007/978-1-0716-1036-7

Kharchenko, P., Tolstorukov, M. & Park, P. Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol 26, 1351–1359 (2008).
https://doi.org/10.1038/nbt.1508

ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3431496/

About motif representation: D'haeseleer, P. What are DNA sequence motifs?. Nat Biotechnol 24, 423–425 (2006). https://doi.org/10.1038/nbt0406-423

General strategies for motif discovery (relatively old paper but gives a good general description of approaches)
https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.0020036

Review of motif discovery algorithms
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6490410/

Pavel Pevzner’s course on bioinformatics algorithms: motif discovery problem
https://youtube.com/playlist?list=PLQ-85IQIPqFMEcdAi0yF015RgmowtsvwT