Multiple Alignments

Nikolai Bykov

Bioinformatics seminar

10.11.2021

Slides are based on materials by Aleksandra Galitsyna

Skoltech

Skolkovo Institute of Science and Technology

Outline

1. Multiple Alignments ~ 30 min
2. Task 1 (part of HW) ~ 30 min
3. Motifs search ~ 45 min
4. Task 2-4 (part of HW) ~ 60 min

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-------ccagccctgg
                **** **  *   ** **    *      **  * *  ***  *  *  * *  **       *     ***
```


slide courtesy of Asya Mendelevich

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:

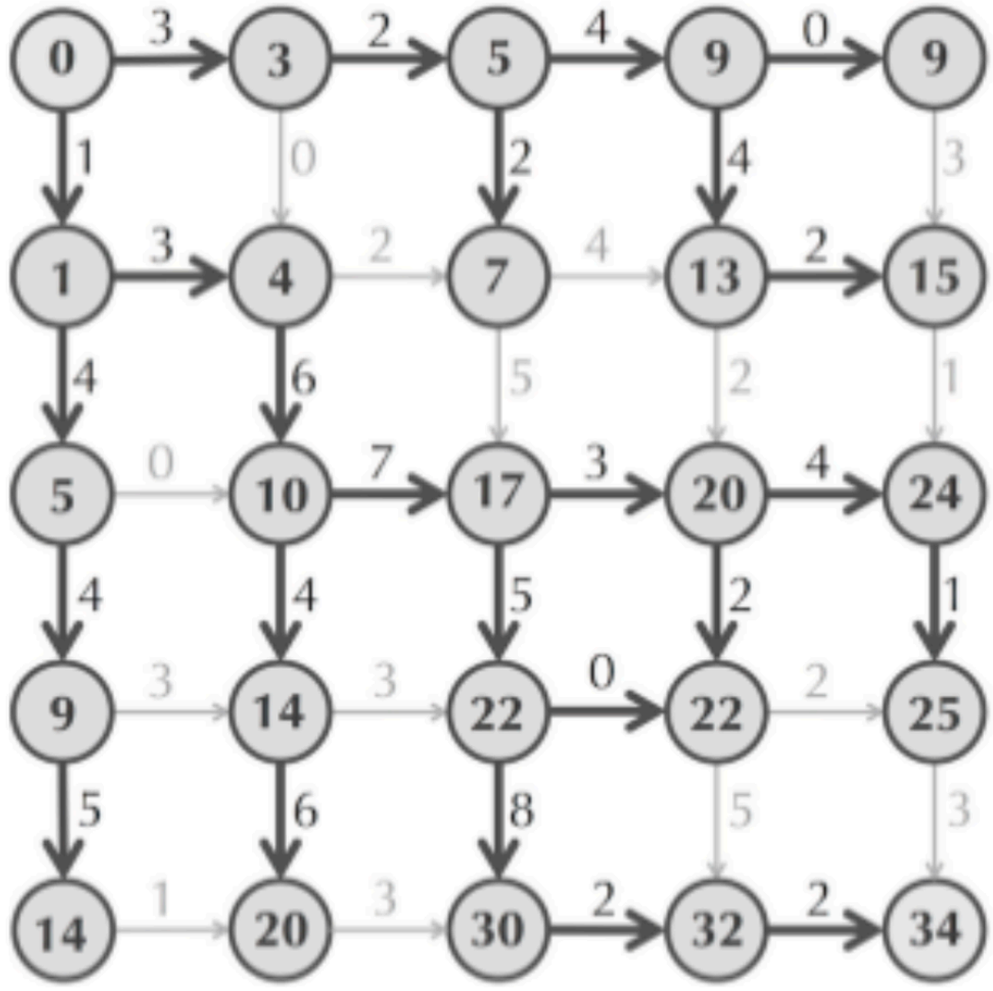
```
>CPZANT
ATGGGAGCGGGGGCGTCTGTTTTGAGGGGAGAGAAGCTAGATACATGGGA
AAGTATCAGGCTTCGGCCCGGTGGCAAGAAAAAGTACATGATAAAACATC
TGGTTTGGGCAAGATCGGAGCTGCAGCGTTTTGCGCTCAGCTCCTCCCTT
CTAGAAACATCAGAAGGTTGTGAAAAGGCTATCCATCAATTGAGCCCTTC
CATAGAAATAAGATCCCCTGAAATAATATCTTTGTTTAACACCATTTGTG
>U455
ATGGGTGCGAGAGCGTCAGTATTAAGCGGGAAAAAATTAGATTCATGGGA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


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

Algorithms for alignment

- Ideal: Dynamical Programming

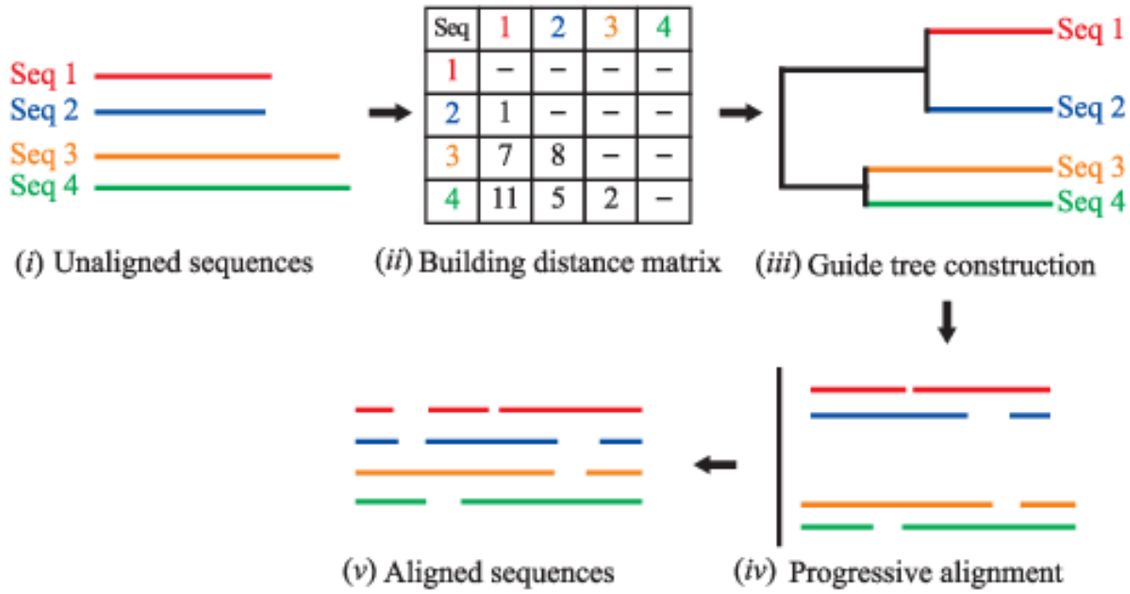
- Heuristics:
  - progressive alignment construction
  - iterative methods
  - consensus methods
  - genetic algorithms



Pavel Pevzner, Phillip Compeau, slide courtesy of Asya Mendelevich

Progressive alignments

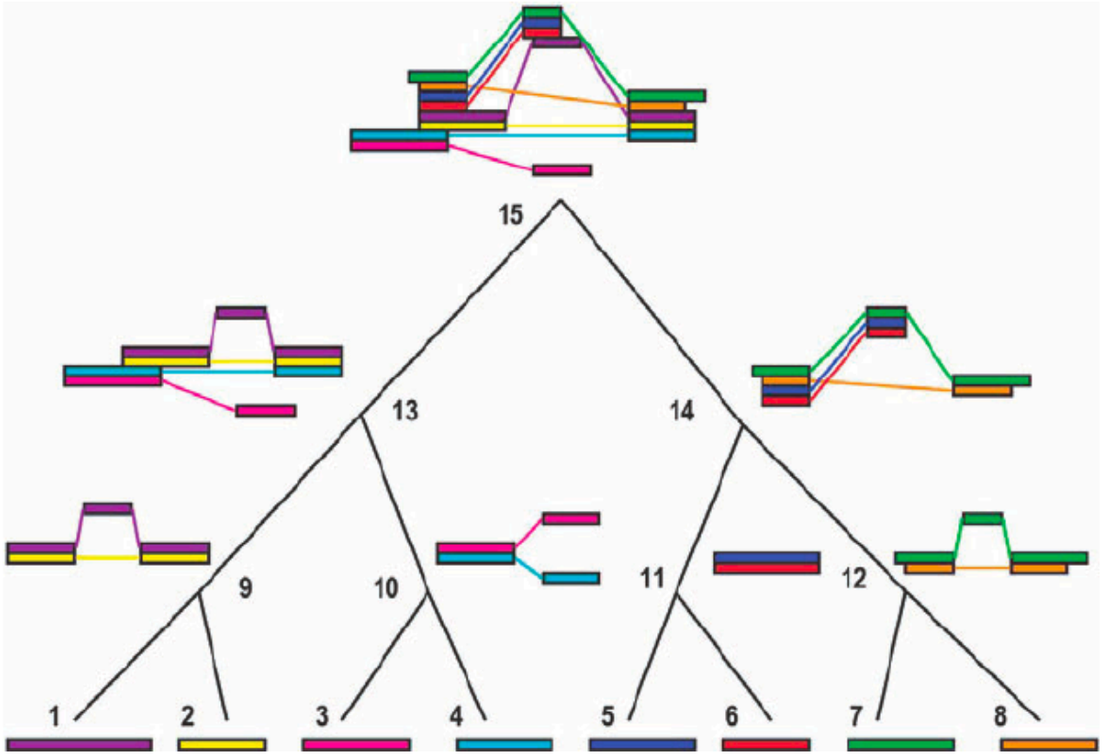
1. Pairwise alignments (each pair of sequences)
2. Builds a distance matrix
3. Finds a guiding tree using one of clustering methods
4. Builds a multiple alignment progressively, starting from most similar sequences, stacking them as in the guiding tree



+ efficient enough to work with up to 1000 sequences

– does not provide a global optimal alignment

– errors in the first steps do propogate 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

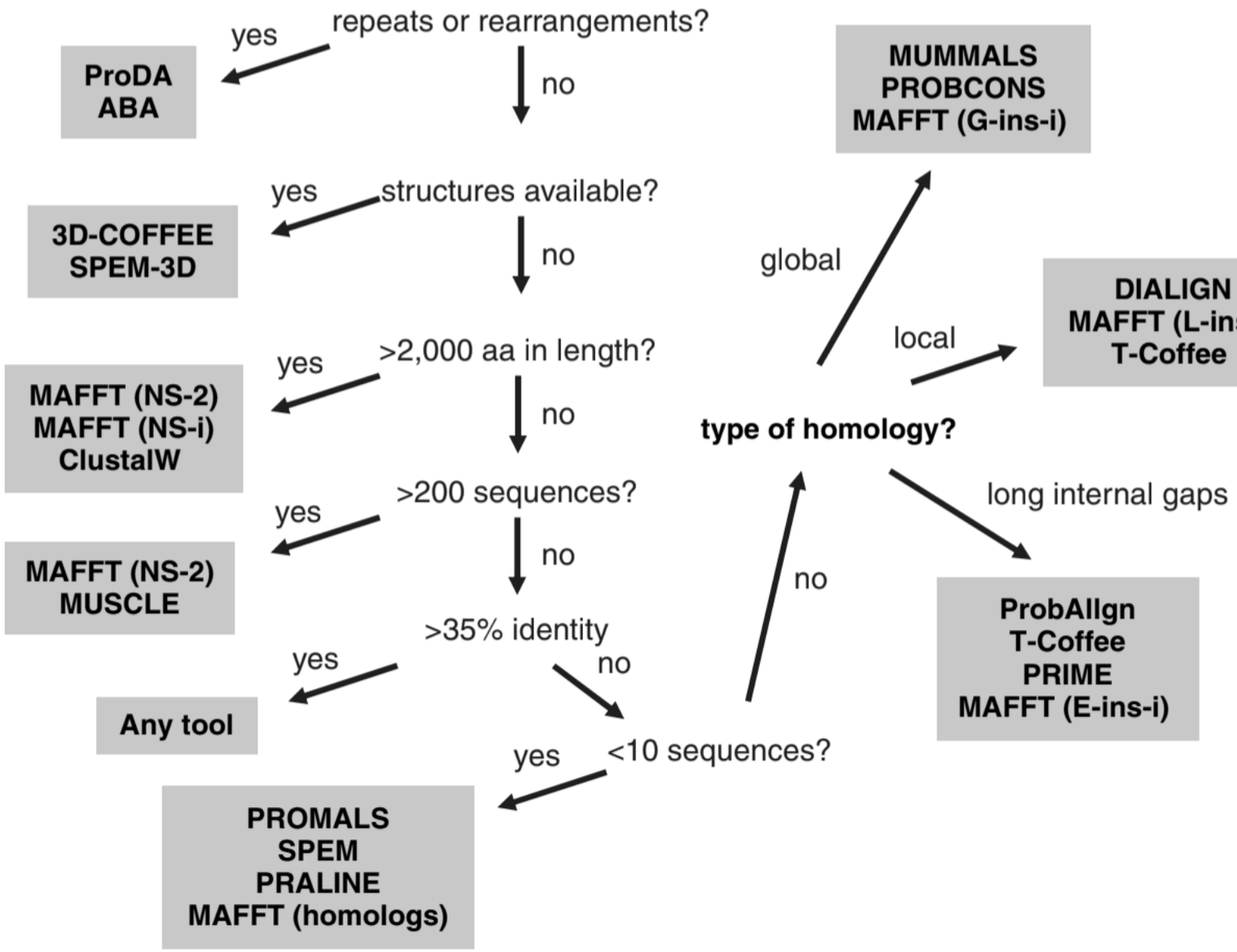
slide courtesy of Asya Mendelevich

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)

https://github.com/encent/2021_Skoltech_Bioinformatics_course_seminar_4#task-1-multiple-alignment

Outline:

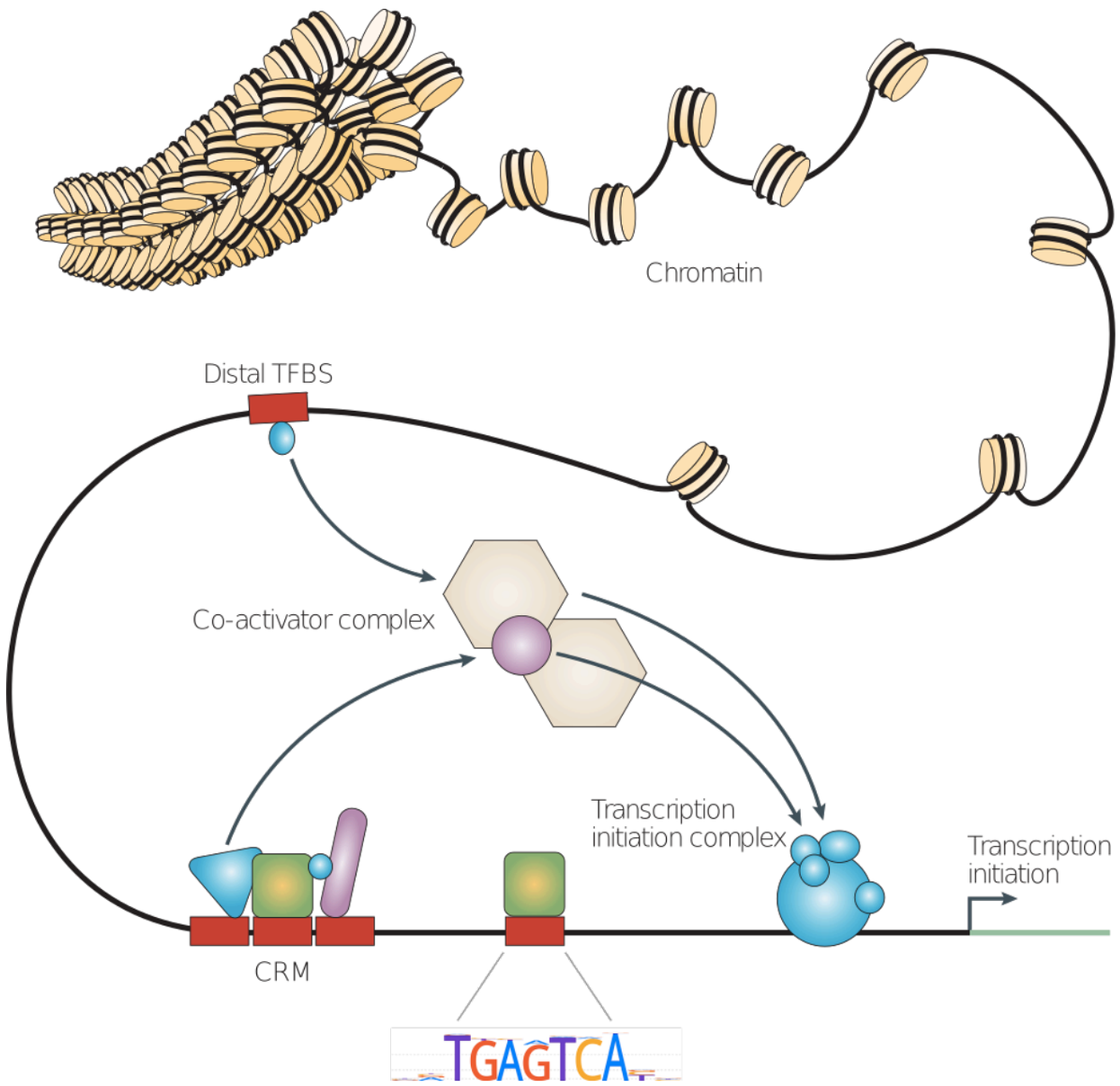
Download file with upstream regions of bacterial orthologs (upstreams.fasta).
Create multiple alignment with T-COFFEE, MUSCLE and CLUSTALW.
Manually select the most conservative gapless region and save it into .fasta file.

We encourage you to run at least **ONE** aligner using **CLI** !

Biological background

Proteins and nucleic acid comprise the main acting components of the cell. Major living processes of the cells are regulated via interactions between them:

- **Protein-DNA:**
  - transcription
  - chromatin modification
  - replication/recombination/ DNA-repair
  - packaging (histones)
- **Protein-RNA:**
  - regulation of RNA metabolism
  - processing: splicing, cleavage and polyadenylation, editing
  - nuclear-cytoplasmic transport
  - translation
  - degradation: microRNA, NMD
- **RNA-DNA interactions:**
  - various chromatin regulation

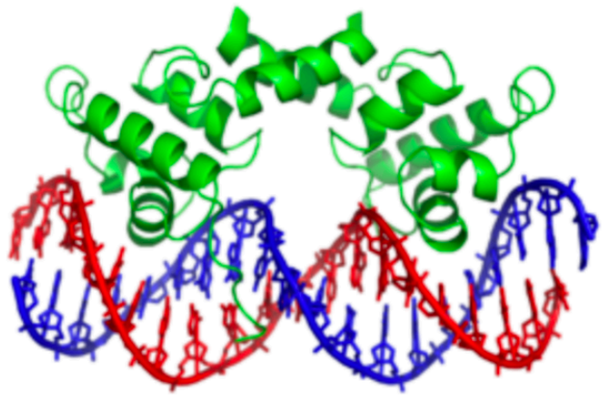


Wasserman & Sandelin, 2004

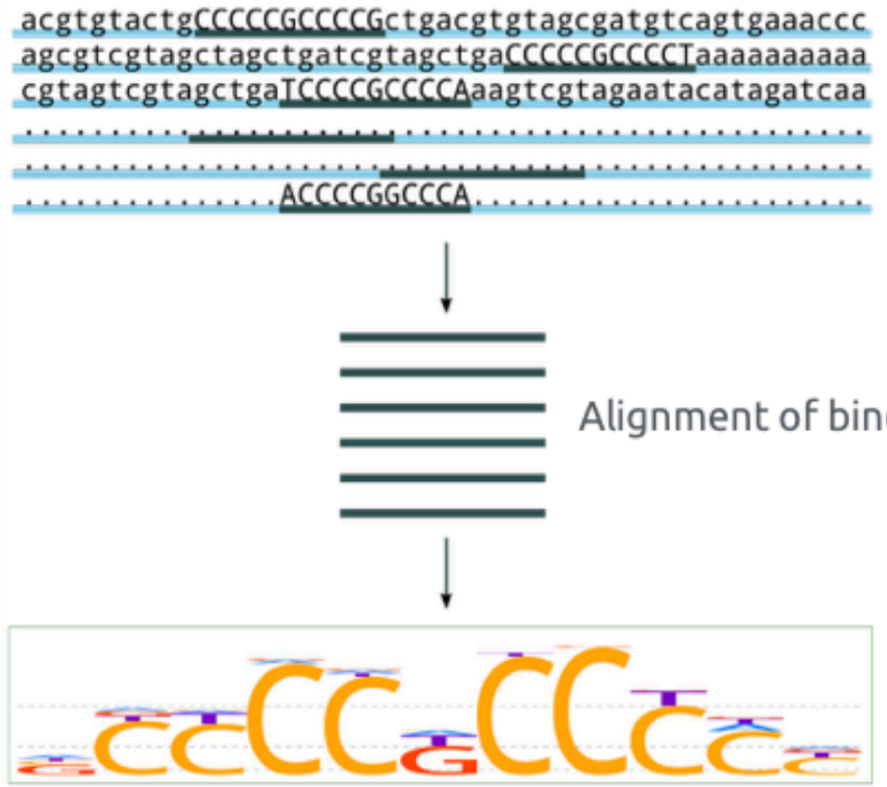
DNA-protein interactions

Binding of protein is typically specific to DNA sequence:

DNA-protein binding



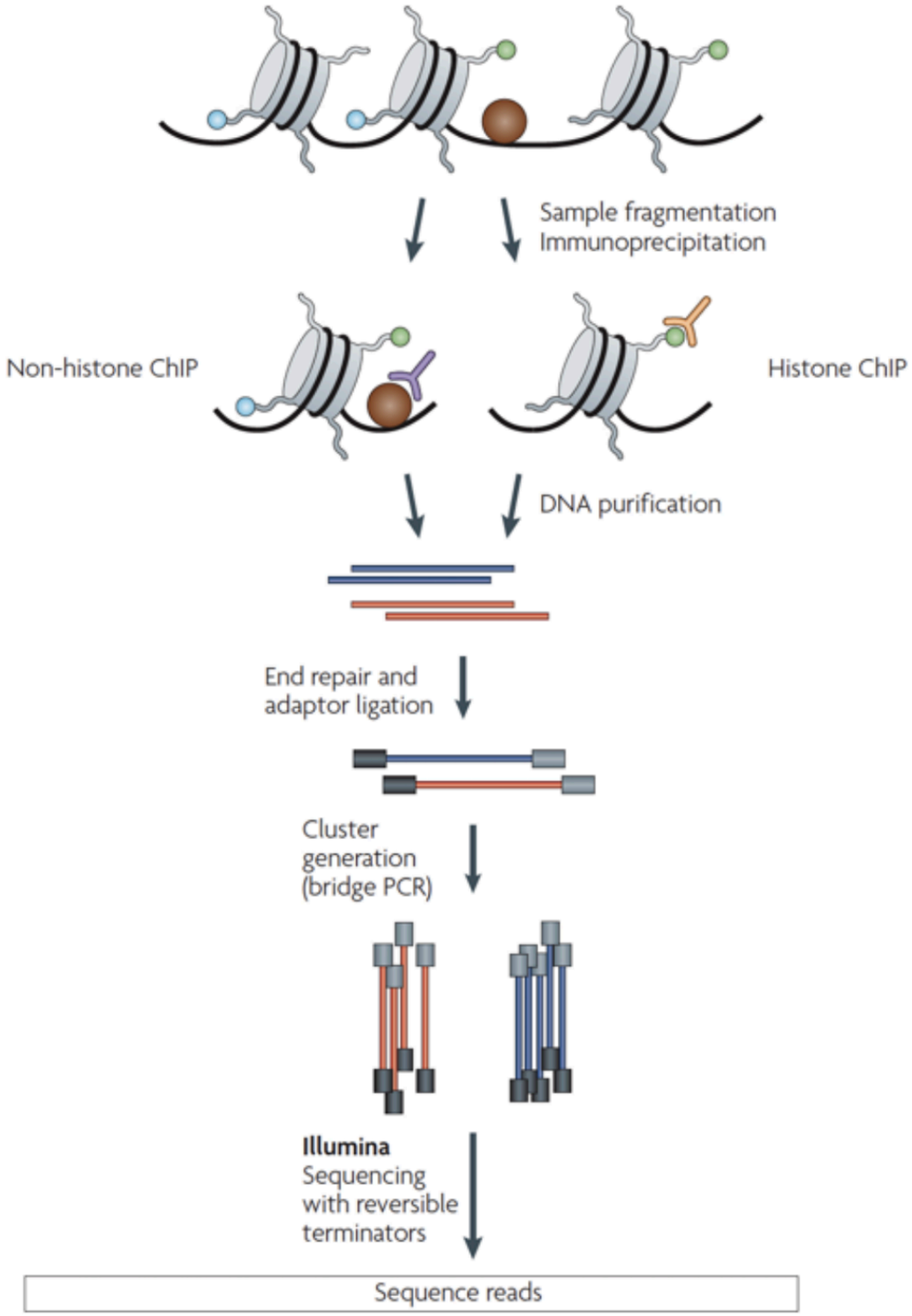
List of DNAs with binding events



Binding pattern

ChIP-Seq basics

Chromatin-immunoprecipitation followed by sequencing:

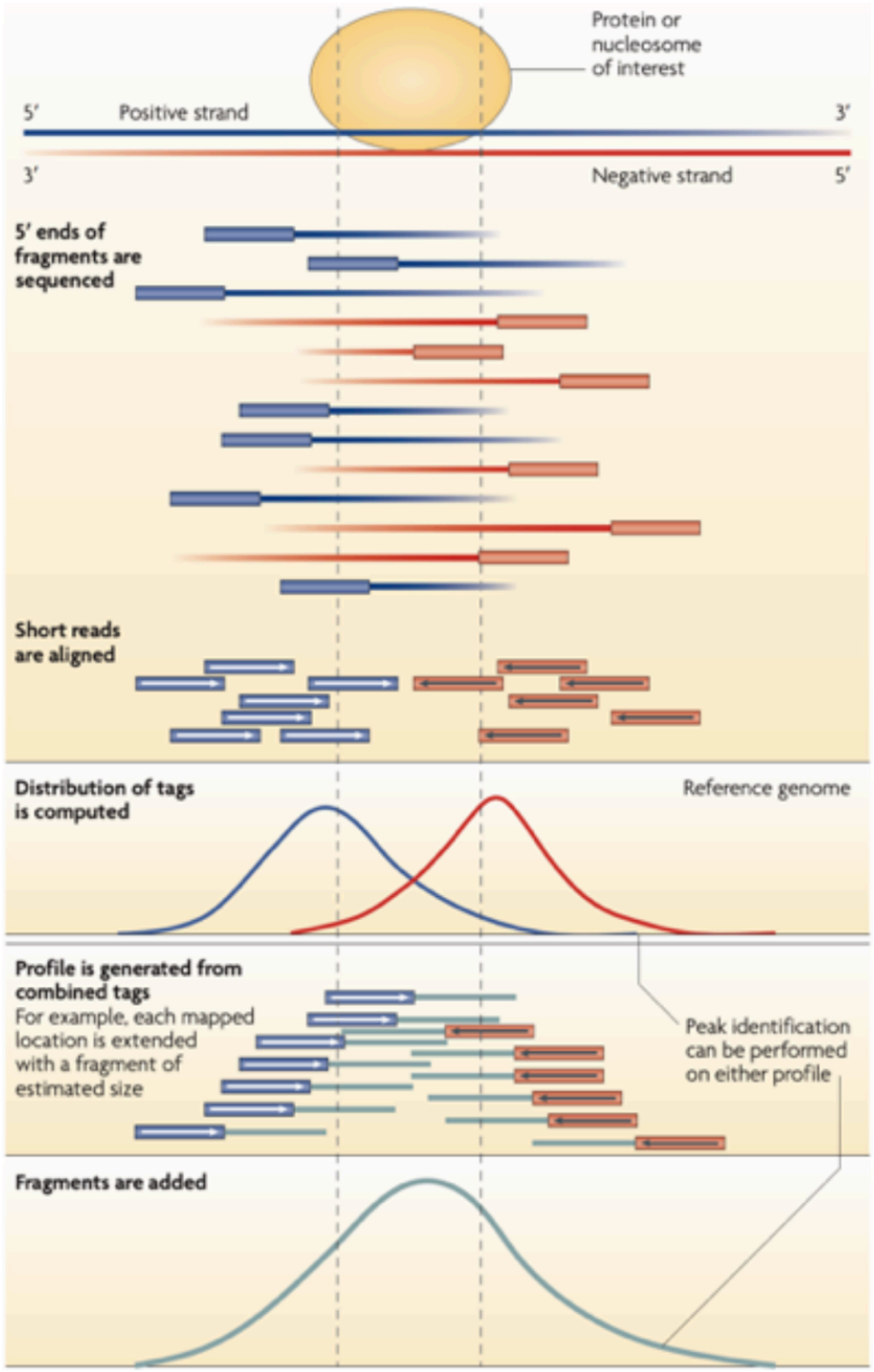


ChIP-Seq basics

Binding events:

Read alignments:

Peak calling:



See you at “Multi-Omics Data Analysis” course at Skoltech

Motifs search: training attention

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

atgaccgggatactgataaaaaaaaggggggggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg

acccctatttttgagcagatttagtgacctggaaaaaaatttgagtacaaaactttccgaataaaaaaaaggggggga

tgagtatccctgggatgacttaaaaaaaaggggggtgctctcccgatttttgaatatgtaggatcattcgccagggtccga

gctgagaattggatgaaaaaaaaggggggtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga

tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaataaaaaaaaggggggcttatag

gtcaatcatgttcttgtgaatggatttaaaaaaaaggggggggaccgcttggcgcacccaaattcagtgtgggcgagcgcaa

cggttttggcccttgttagaggcccccgtaaaaaaaaggggggcaattatgagagagctaatctatcgcgtgcgtgttcat

aacttgagttaaaaaaaaggggggctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta

ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcataaaaaaaaggggggggaccgaaagggaag

ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttaaaaaaaaggggggga

Motifs search: training attention

- Seems to be easy:

atgaccgggatactgatAAAAAAAAGGGGGGGggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg

acccctatttttgagcagatttagtgacctggaaaaaaatttgagtacaaaacttttccgaataAAAAAAAAGGGGGGGa

tgagtatccctgggatgacttAAAAAAAAGGGGGGGtgctctcccgattttgaatatgtaggatcattcgccagggtccga

gctgagaattggatgAAAAAAAAGGGGGGGtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga

tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaatAAAAAAAAGGGGGGGcttatag

gtcaatcatgttcttgtgaatggatttAAAAAAAAGGGGGGGgaccgcttggcgcacccaaattcagtgtgggcgagcgcaa

cggttttggcccttgttagaggccccgtAAAAAAAAGGGGGGGcaattatgagagagctaatctatcgcgtgcgtgttcat

aacttgagttAAAAAAAAGGGGGGGctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta

ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatAAAAAAAAGGGGGGGaccgaaagggaag

ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttAAAAAAAAGGGGGGGa

Motifs search: training attention

- Let's introduce some substitutions:

atgaccgggatactgatAgAAgAAAGGttGGGggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg

acccctatttttgagcagatttagtgacctggaaaaaaatttgagtacaaaacttttccgaatacAAtAAAAcGGcGGGa

tgagtatccctgggatgacttAAAAtAAtGGaGtGGtgctctcccgatttttgaatatgtaggatcattcgccagggtccga

gctgagaattggatgcAAAAAAAGGGattGtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga

tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaatAtAAtAAAGGaaGGGcttatag

gtcaatcatgttcttgtgaatggatttAAcAAtAAGGGctGGgaccgcttggcgcacccaaattcagtgtgggcgagcgcaa

cggttttggcccttgttagaggcccccgtAtAAAcAAGGaGGGccaattatgagagagctaatctatcgcgtgcgtgttcat

aacttgagttAAAAAAtAGGGaGccctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta

ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatActAAAAAGGaGcGGaccgaaagggaag

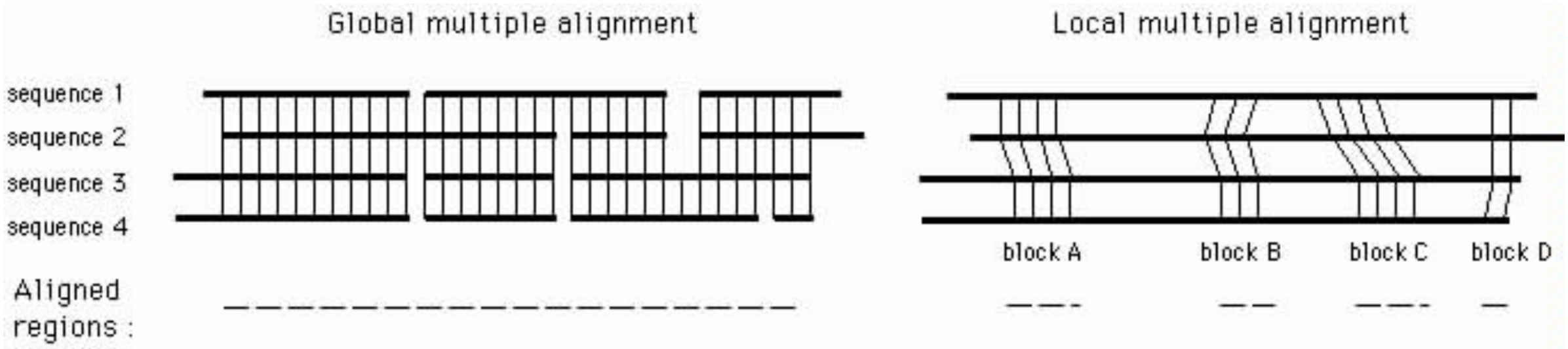
ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttActAAAAAGGaGcGGa

Motifs search: training attention

- Is everything easy if you know the answer?

```
atgaccgggatactgatagaagaaaggttgggggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg
acccctatttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaacttttccgaatacaataaaacggcggga
tgagtatccctgggatgacttaaaataatggagtggtgctctcccgatttttgaatatgtaggatcattcgccagggtccga
gctgagaattggatgcaaaaaaagggattgtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga
tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaatataataaaggaagggcttatag
gtcaatcatgttcttgtgaatggatttaacaataagggctgggaccgcttggcgcacccaaattcagtgtgggcgagcgcaa
cggttttggcccttgttagaggcccccgtataaacaaggagggccaattatgagagagctaatctatcgcgtgcgtgttcat
aacttgagttaaaaaatagggagccctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta
ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatactaaaaaggagcggaccgaaagggaag
ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttactaaaaaggagcgga
```

Multiple alignment: global vs local



Motif representation: consensus sequence

Consensus sequence lists nucleotides that are allowed in given position.
Consider following gapless block of local 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: position weight matrix

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

frequency matrix

	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

probability matrix

	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

b_n is probability of nucleotide n in background

	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

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

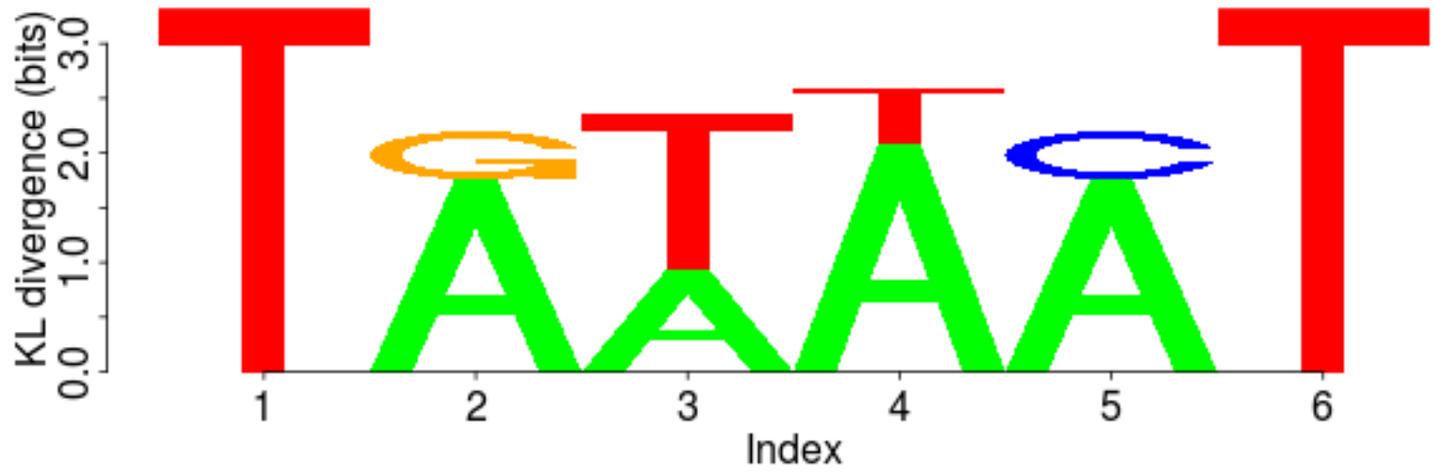
	1	2	3	4	5	6
A	-1.2	1.2	0.4	1.2	1.2	-1.2
C	-1.2	-1.2	-1.2	-1.2	-0.2	-1.2
G	-1.2	-0.2	-1.2	-1.2	-1.2	-1.2
T	1.4	-1.2	0.8	-0.2	-1.2	1.4

PWM visualisation: logo

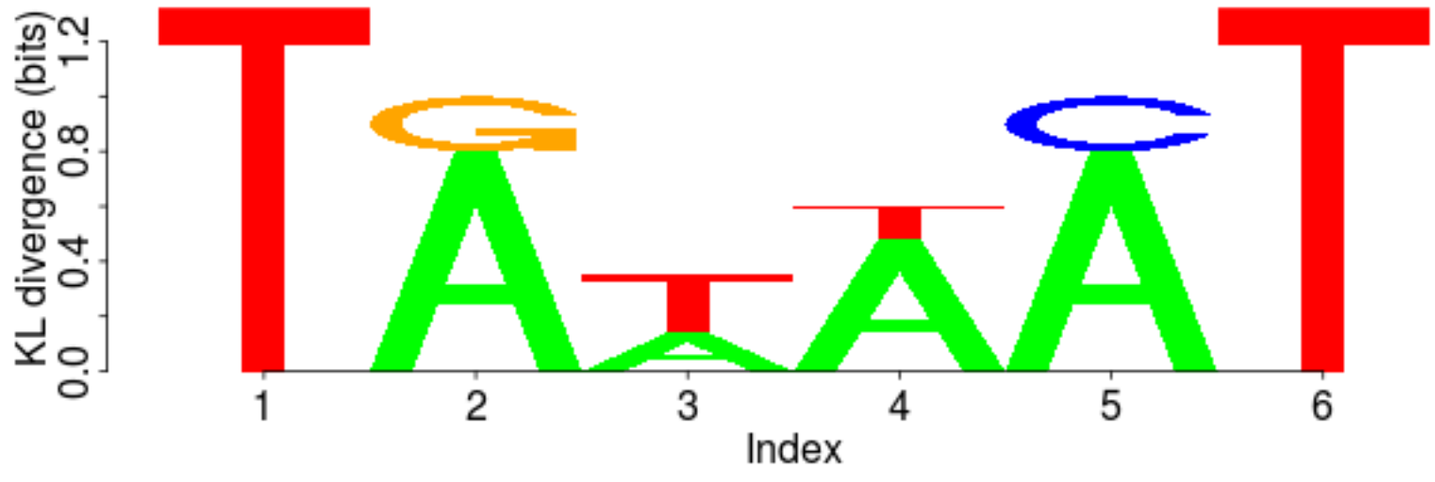
Height of each column is significance of the position (dissimilarity to background).

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

GC-rich background:



AT-rich background:



Search of motif instances in a given sequence

- 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

T G A T A C

given sequence (e.g. genome)

	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

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

score

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

- Additivity assumption: score is larger for longer sequences!

Search of motif instances

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

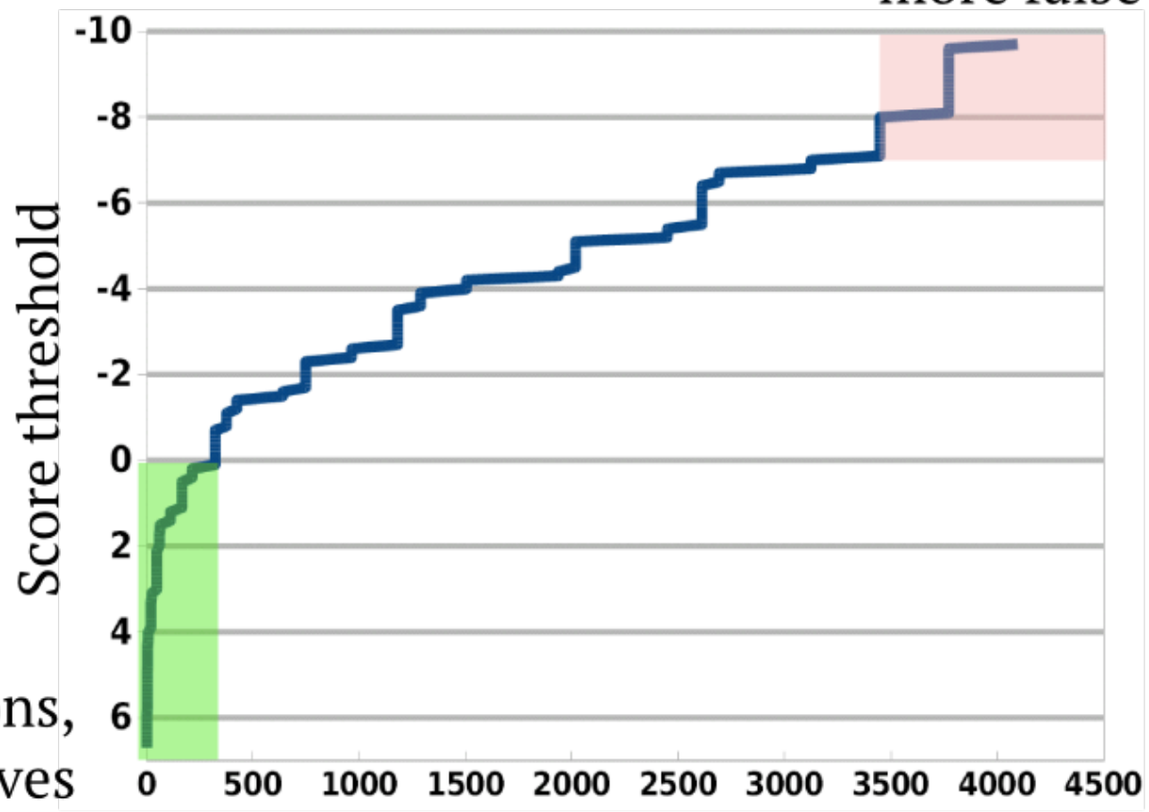
PWM

GGATTA → S_{GGATTA}=1.22+1.22+0.96+1.22+1.22+0.96=**6.8** the best score

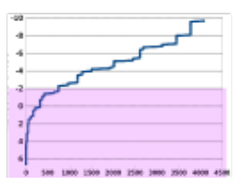
S_{GGGGGG}=2.44-6.4=**-3.96**

S=**-9.6** the worst score

more predicted TFBS,
more false positive predictions



less TFBS predictions,
less true positives



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

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 (Motifs search)

All the materials for this seminar are located on GitHub:

https://github.com/encent/2021_Skoltech_Bioinformatics_course_seminar_4

- 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).