

# Множественное выравнивание

## Поиск мотивов

Практическая биоинформатика 08.10.2024

Мичил Трофимов

Материалы взяты и адаптированы из лекции **Анны Рыбиной** «Поиск мотивов» в рамках курса «Биоинформатика» в магистратуре Сколтеха «Науки о жизни»



1. Мотив, проблема обнаружения мотивов ~ 15
2. Множественное выравнивание последовательностей ~ 20
3. Задание 1 (часть HW) ~ 30
4. ChIP-seq ~ 10
5. Представление мотивов ~ 20
6. Задачи 2-4 (часть HW) ~ 45
7. Сканирование мотивов ~ 10
8. Обучение командной строке в классе ~ 10
9. Задание 5 (часть HW)

# Мотив

**Мотив** - это повторяющийся (консервативный) паттерн, который, как предполагается, имеет биологическое значение (обладает биологической функцией)

может встречаться в:

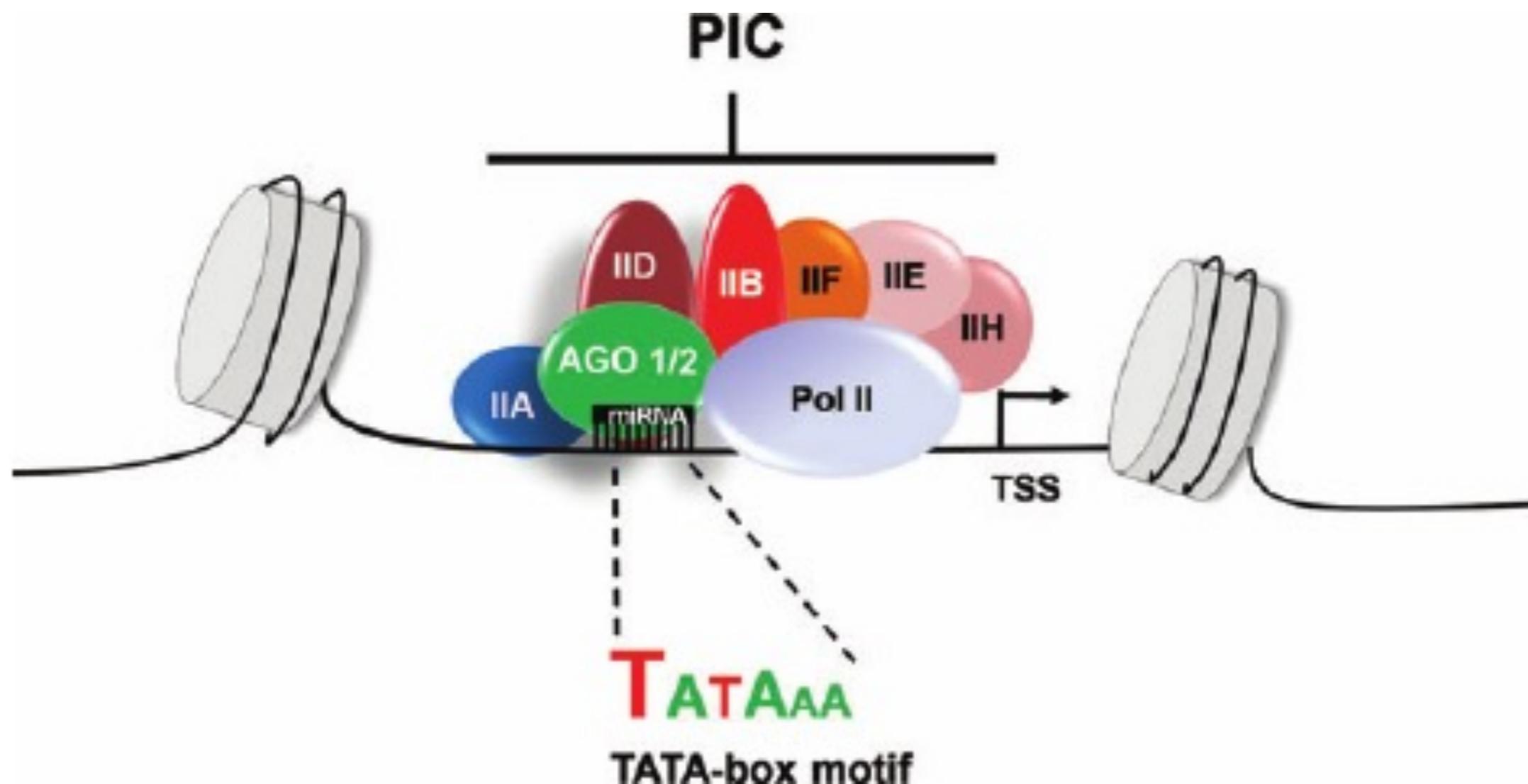
- структуре или последовательности
- РНК/белок/ДНК

может быть вовлечен во взаимодействие с другими молекулами (белками/нуклеиновыми кислотами)

Основные процессы жизнедеятельности клеток регулируются через взаимодействие между белками и нуклеиновыми кислотами: белок-ДНК, белок-РНК, РНК-ДНК

На семинаре мы рассмотрим мотивы последовательностей ДНК, распознаваемые белком -- транскрипционным фактором

# Мотив последовательности ДНК



# Как найти мотив в ДНК?

## Экспериментальный подход:

- ДНК-следы
- SELEX
- анализы сдвига электрофоретической подвижности

## Примеры:

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3080775/>

## Вычислительный подход - задача поиска мотивов:

- поиск перепредставленных (и/или консервативных) паттернов ДНК перед функционально связанными генами (например, генами со схожими паттернами экспрессии)

# Задача поиска мотивов

Задача поиска мотивов может быть сформулирована следующим образом:

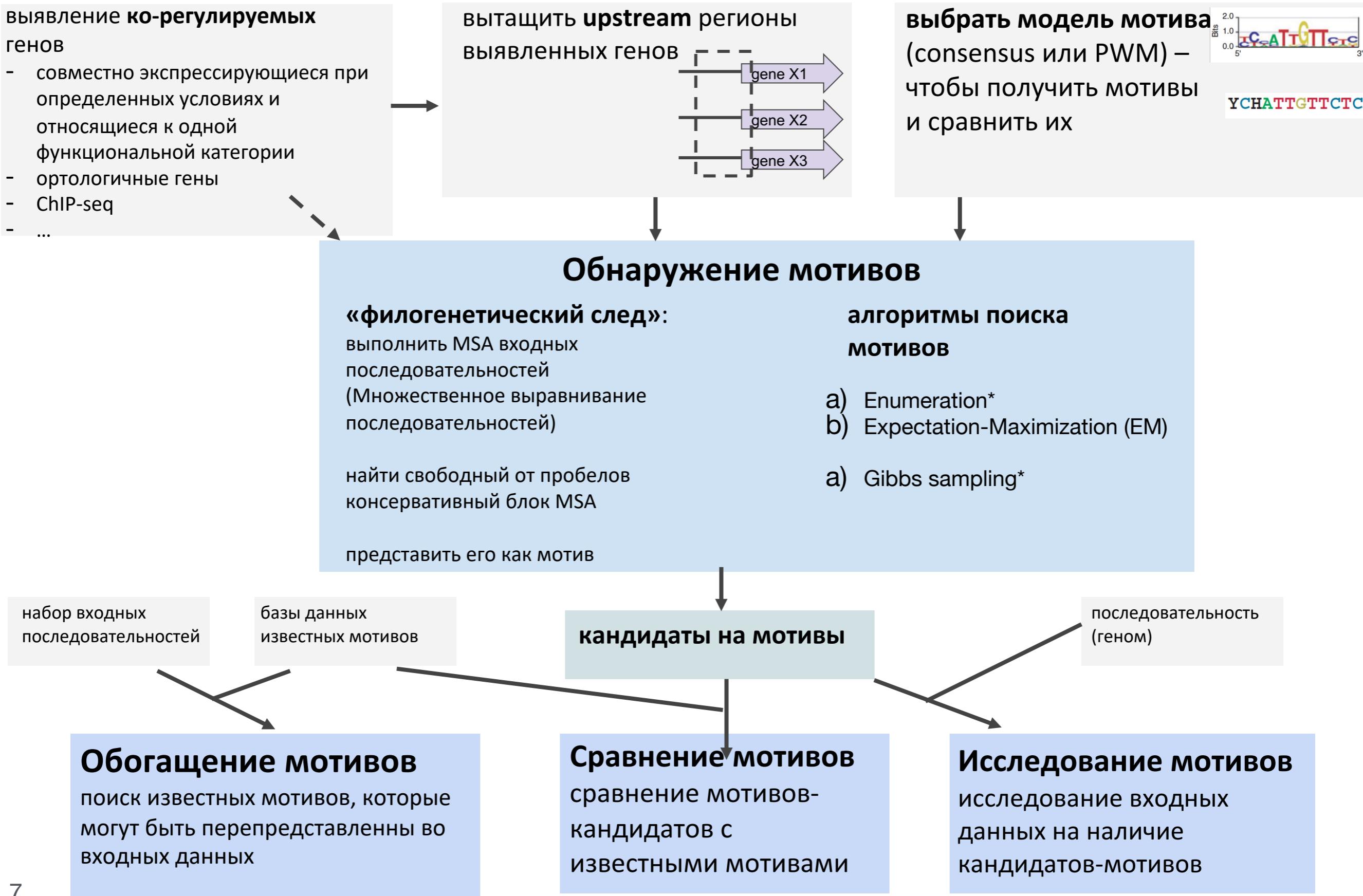
**Дано:** набор последовательностей ДНК

**Допущение:** соответствующие гены **корегуляторные** и, следовательно, могут быть связаны одним или несколькими регуляторными белками

**Найти:** параметры мотива(ов), которые могут объяснить это связывание:

- количество мотивов
- ширина каждого мотива
- его расположение в исходных последовательностях

# Задача поиска мотивов: flowchart



# Задача поиска мотивов: flowchart

**выявление ко-регулируемых генов**

- совместно экспрессирующиеся при определенных условиях и относящиеся к одной функциональной категории
- ортологичные гены
- ChIP-seq
- ...

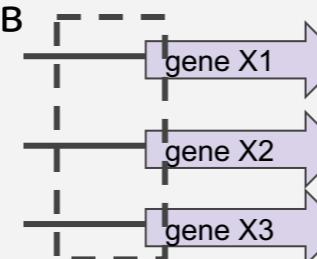
# Задача поиска мотивов: flowchart

## выявление ко-регулируемых генов

- совместно экспрессирующиеся при определенных условиях и относящиеся к одной функциональной категории
- ортологичные гены
- ChIP-seq
- ...

## вытащить **upstream** регионы

### выявленных генов

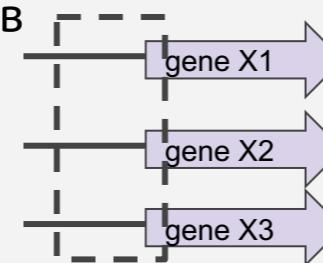


# Задача поиска мотивов: flowchart

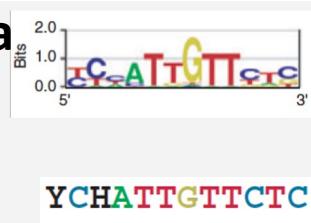
## выявление ко-регулируемых генов

- совместно экспрессирующиеся при определенных условиях и относящиеся к одной функциональной категории
- ортологичные гены
- ChIP-seq
- ...

## вытащить **upstream** регионы выявленных генов



**выбрать модель мотива** (consensus или PWM) – чтобы получить мотивы и сравнить их

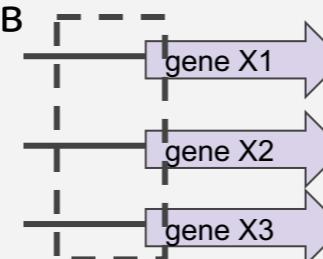


# Задача поиска мотивов: flowchart

выявление ко-регулируемых генов

- совместно экспрессирующиеся при определенных условиях и относящиеся к одной функциональной категории
- ортологичные гены
- ChIP-seq
- ...

вытащить **upstream** регионы выявленных генов



выбрать модель мотива (consensus или PWM) – чтобы получить мотивы и сравнить их



GCATTCGTTCTC

## Обнаружение мотивов

«филогенетический след»:

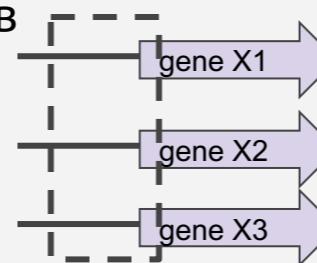
алгоритмы поиска мотивов

# Задача поиска мотивов: flowchart

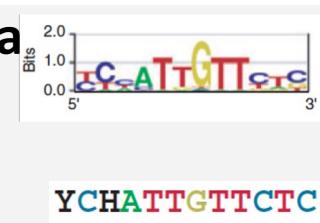
## выявление ко-регулируемых генов

- совместно экспрессирующиеся при определенных условиях и относящиеся к одной функциональной категории
- ортологичные гены
- ChIP-seq
- ...

## вытащить **upstream** регионы выявленных генов



## выбрать модель мотива (consensus или PWM) – чтобы получить мотивы и сравнить их



## Обнаружение мотивов

### «филогенетический след»:

выполнить MSA входных последовательностей  
(Множественное выравнивание последовательностей)

найти свободный от пробелов консервативный блок MSA

представить его как мотив

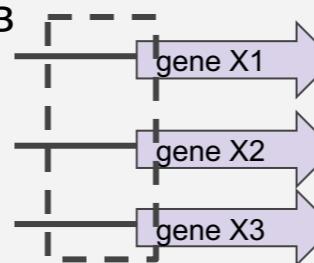
### алгоритмы поиска мотивов

# Задача поиска мотивов: flowchart

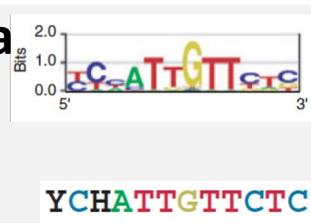
## выявление ко-регулируемых генов

- совместно экспрессирующиеся при определенных условиях и относящиеся к одной функциональной категории
- ортологичные гены
- ChIP-seq
- ...

## вытащить **upstream** регионы выявленных генов



## выбрать модель мотива (consensus или PWM) – чтобы получить мотивы и сравнить их



## Обнаружение мотивов

### «филогенетический след»:

выполнить MSA входных последовательностей  
(Множественное выравнивание последовательностей)

найти свободный от пробелов консервативный блок MSA

представить его как мотив

### алгоритмы поиска мотивов

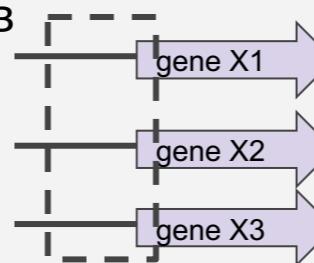
- a) Enumeration\*
- b) Expectation-Maximization (EM)
- a) Gibbs sampling\*

# Задача поиска мотивов: flowchart

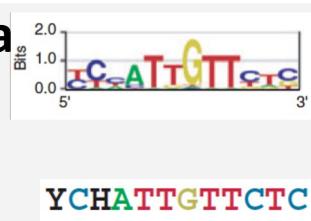
## выявление ко-регулируемых генов

- совместно экспрессирующиеся при определенных условиях и относящиеся к одной функциональной категории
- ортологичные гены
- ChIP-seq
- ...

## вытащить **upstream** регионы выявленных генов



## выбрать модель мотива (consensus или PWM) – чтобы получить мотивы и сравнить их



## Обнаружение мотивов

### «филогенетический след»:

выполнить MSA входных последовательностей  
(Множественное выравнивание последовательностей)

найти свободный от пробелов консервативный блок MSA

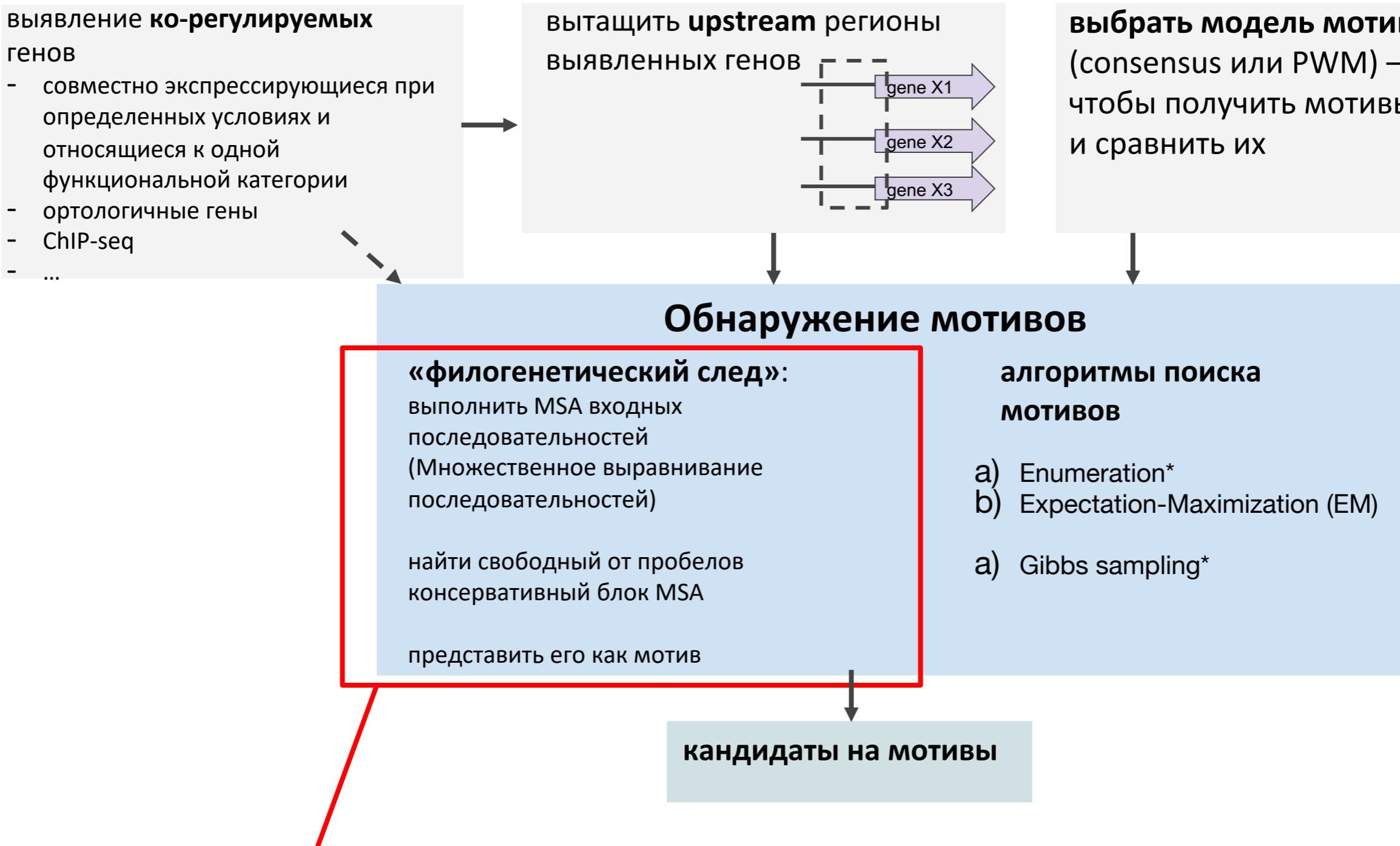
представить его как мотив

### алгоритмы поиска мотивов

- a) Enumeration\*
- b) Expectation-Maximization (EM)
- a) Gibbs sampling\*

кандидаты на мотивы

# Задача поиска мотивов: flowchart



Y C H A T T G T T C T C

Как использовать множественное выравнивание для определения мотивов (TASK 1)

# Множественное выравнивание

- Может быть применимо к любой последовательности (ДНК, РНК, белки или другое)
  - Парное выравнивание (2 последовательности)

- Множественное выравнивание( $\geq 3$  последовательностей):

# Множественное выравнивание: применение

- **филогенетический анализ**
- **структурная биоинформатика**
- **поиск мотивов**

# Множественное выравнивание: применение

- **филогенетический анализ**
- **структурная биоинформатика**
- **поиск мотивов**

# Форматы файлов выравнивания

## Clustal W:

```
CPZANT ATGGGAGCGGGGGCGTCTGTTTGAGGGGAGAGAAGCTAGATAACATGGGA
U455   ATGGGTGCGAGAGCGTCAGTATTAAGCGGGAAAAAAATTAGATTACATGGGA

CPZANT AAGTATCAGGCTTCGGCCCGTGGCAAGAAAAAGTACATGATAAAACATC
U455   GAAAATT CGGTTAACGCCAGGGGGAAACAAAAAATATAGACTGAAACATT

CPZANT TGGTTTGGGCAAGATCGGAGCTGCAGCGTTGCGCTCAGCTCCCTCCCTT
U455   TAGTATGGGCAAGCAGGGAGCTGGAAAAATTCAACTAACCTGGCCTT

CPZANT CTAGAAACATCAGAAGGTTGTGAAAAGGCTATCCATCAATTGAGCCCTTC
U455   TTAGAAACAGCAGAAGGATGTCAGCAAATACTGGGACAATTACAACCAGC

CPZANT CATAGAAATAAGATCCCCTGAAATAATATCTTGTAAACACCATTGTG
U455   TCTCCAGACAGGAACAGAAGAACTTAGATCATTATATAATACAGTAGCAG
```

## FastA:

```
>CPZANT
ATGGGAGCGGGGGCGTCTGTTTGAGGGGAGAGAAGCTAGATAACATGGGA
AAGTATCAGGCTTCGGCCCGTGGCAAGAAAAAGTACATGATAAAACATC
TGGTTTGGGCAAGATCGGAGCTGCAGCGTTGCGCTCAGCTCCCTCCCTT
CTAGAAACATCAGAAGGTTGTGAAAAGGCTATCCATCAATTGAGCCCTTC
CATAGAAATAAGATCCCCTGAAATAATATCTTGTAAACACCATTGTG
>U455
ATGGGTGCGAGAGCGTCAGTATTAAGCGGGAAAAAAATTAGATTACATGGGA
GAAAATT CGGTTAACGCCAGGGGGAAACAAAAAATATAGACTGAAACATT
TAGTATGGGCAAGCAGGGAGCTGGAAAAATTCAACTAACCTGGCCTT
TTAGAAACAGCAGAAGGATGTCAGCAAATACTGGGACAATTACAACCAGC
TCTCCAGACAGGAACAGAAGAACTTAGATCATTATATAATACAGTAGCAG
```

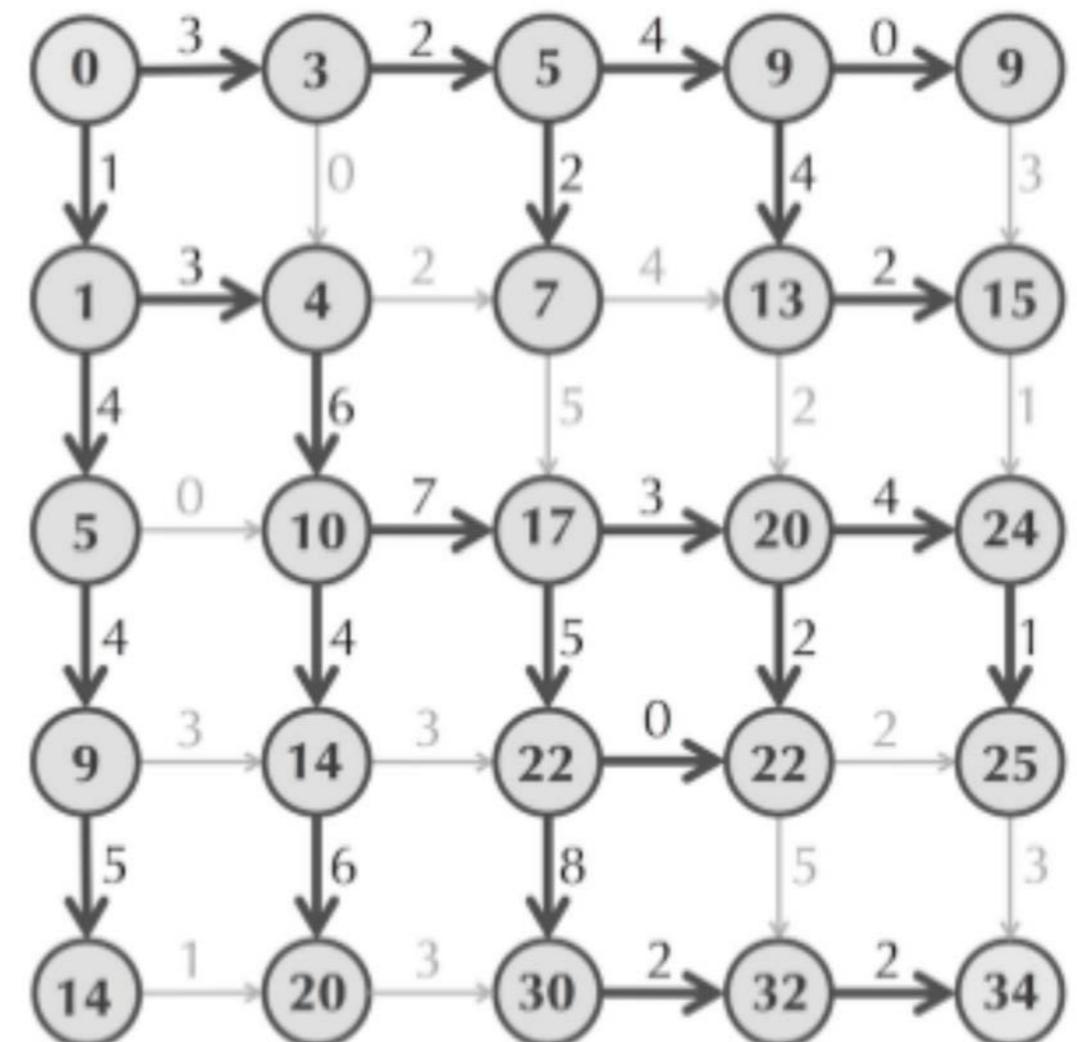
More examples:

<https://www.hiv.lanl.gov/content/sequence/HelpDocs/SEQsamples.html>

# Алгоритмы для множественного выравнивания

- Идеальный вариант: **Динамическое программирование** - оптимальное решение, но очень сложно подсчитать
- **Эвристики\*** - подход, позволяющий уменьшить сложность задачи (~ сделать вычисления быстрее):
  - прогрессивное выравнивание
  - итеративные методы
  - консенсусные методы
  - генетические алгоритмы

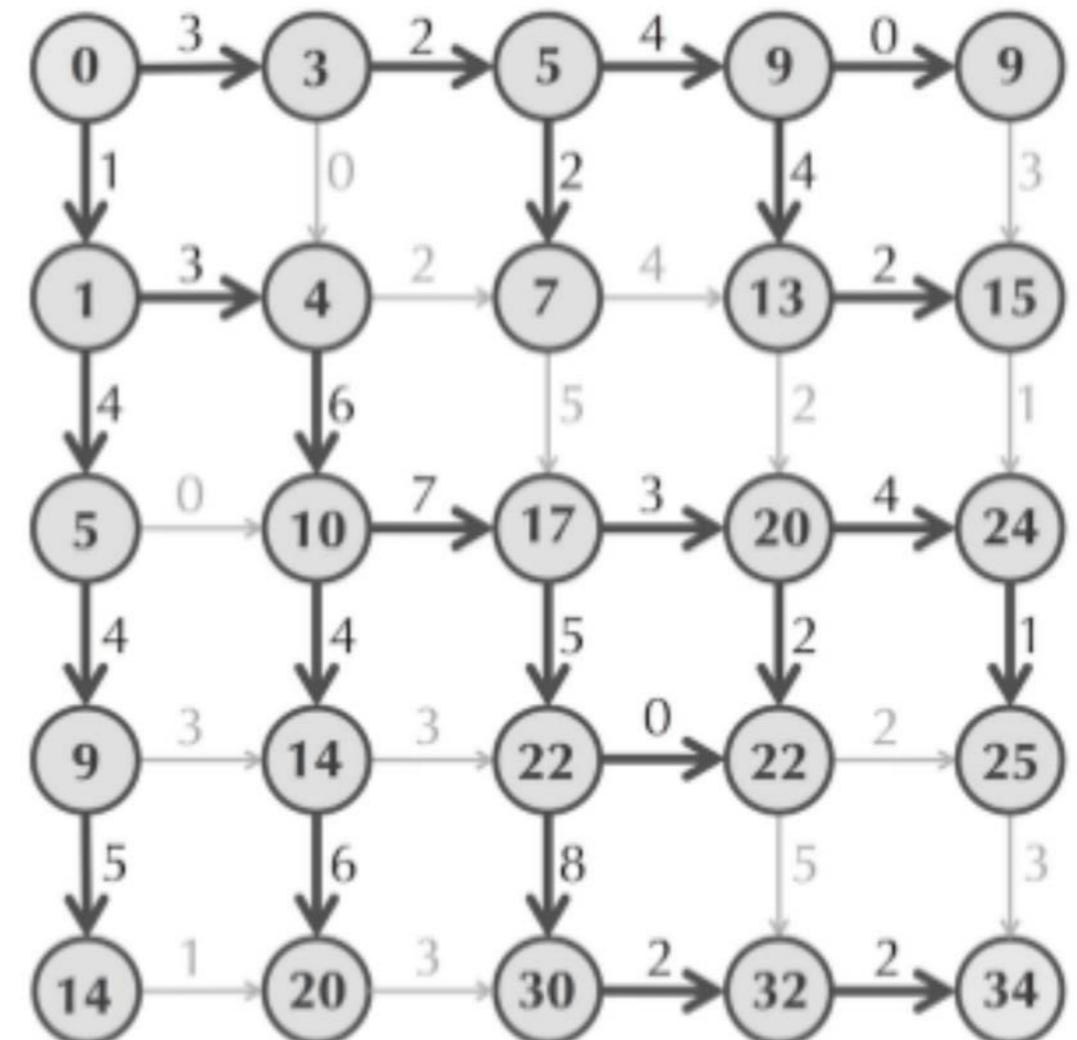
\*Эвристика - это алгоритм, который способен дать приемлемое решение проблемы во многих практических сценариях, но для которого не существует формального доказательства его правильности



# Алгоритмы для множественного выравнивания

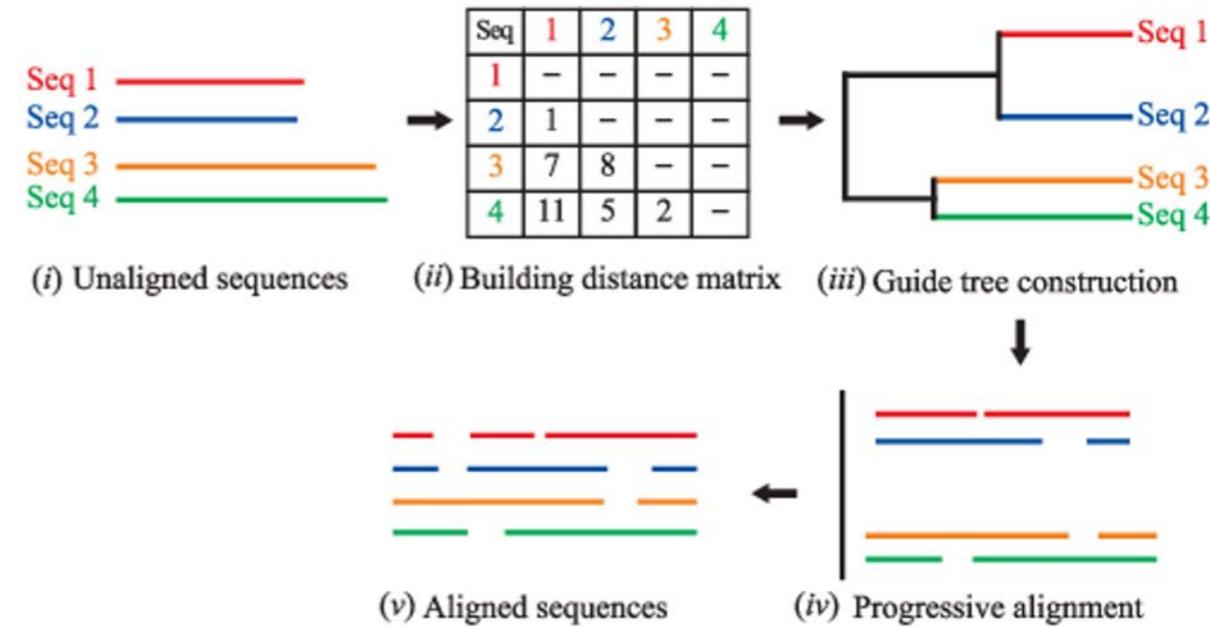
- Идеальный вариант: **Динамическое программирование** - оптимальное решение, но очень сложно подсчитать
- **Эвристики\*** - подход, позволяющий уменьшить сложность задачи (~ сделать вычисления быстрее):
  - **прогрессивное выравнивание**
  - **итеративные методы**
  - **консенсусные методы**
  - **генетические алгоритмы**

\*Эвристика - это алгоритм, который способен дать приемлемое решение проблемы во многих практических сценариях, но для которого не существует формального доказательства его правильности

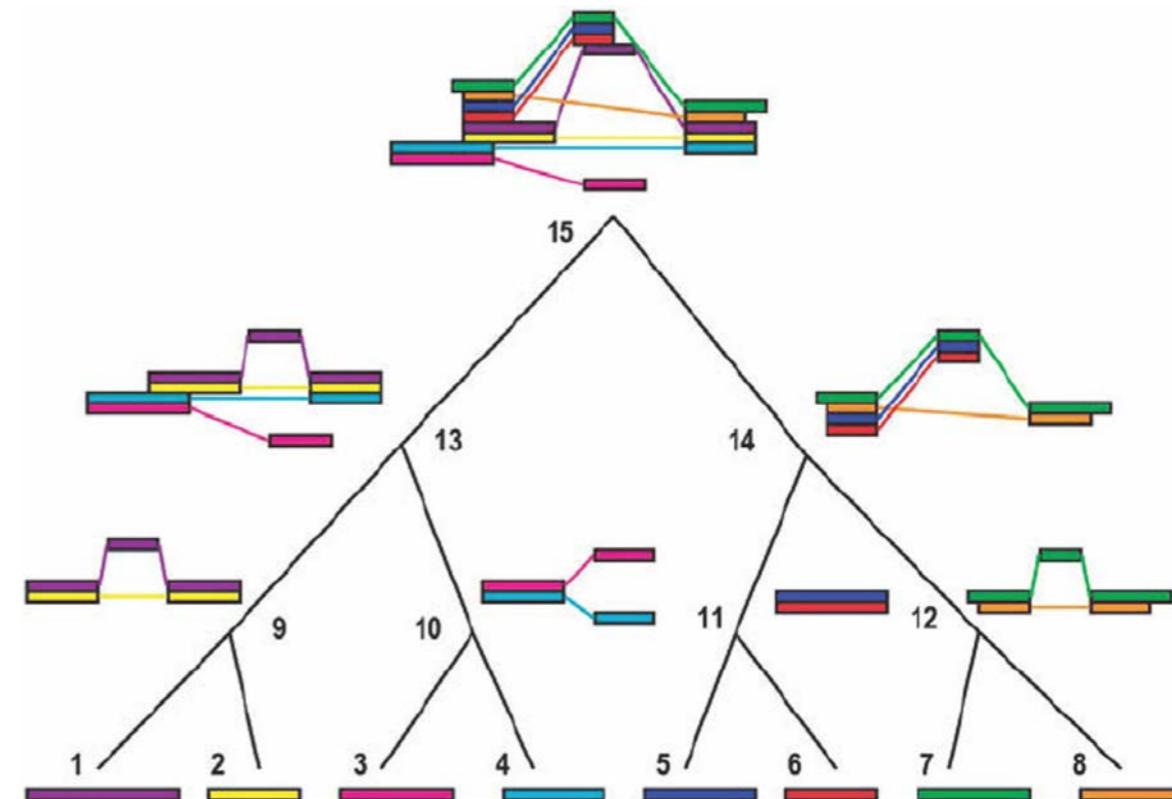


# Прогрессивное выравнивание

1. Попарное выравнивание (каждая пара последовательностей)
2. Построение матрицы расстояний
3. Находит направляющее дерево, используя один из методов кластеризации
4. Постепенно строит множественное выравнивание, начиная с наиболее похожих последовательностей, укладывая их в направляющее дерево



- + достаточно эффективен для работы с 1000 последовательностей
- не обеспечивает глобального оптимального выравнивания
- ошибки на первых этапах (например, ошибочные пропуски) распространяются на окончательное выравнивание



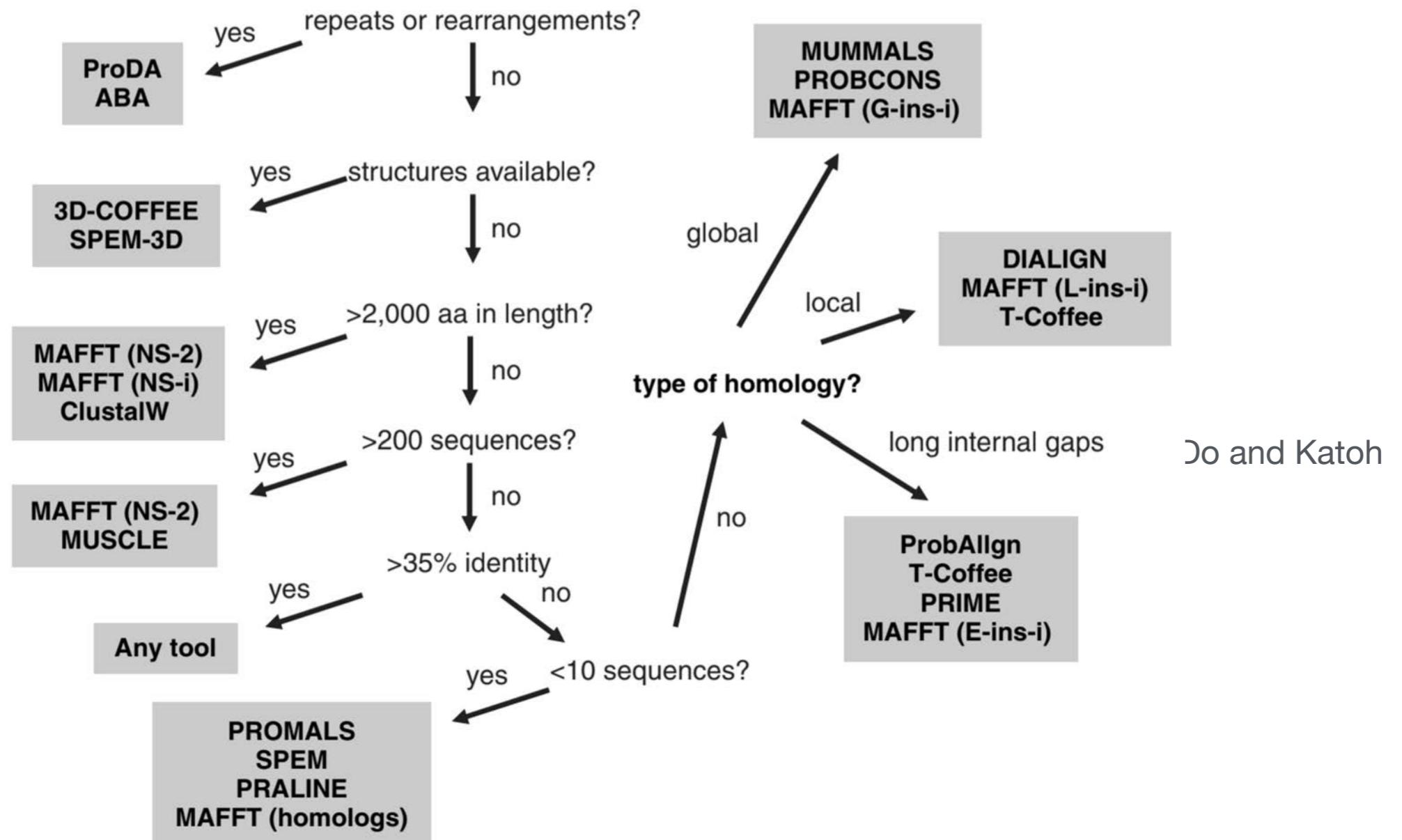
# Итеративные методы

- работает аналогично прогрессивным алгоритмам, но позволяет перестраивать последовательности в выравнивании на каждом шаге
- оптимизирует глобальную метрику
  - + менее подвержен распространению ошибок, обеспечивает более точный результат
  - + хорошо работает с попарно удаленными последовательностями
  - всё ещё эвристика
  - не так эффективен, как прогрессивные алгоритмы

# Популярные тулы для множественного выравнивания

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

# Как выбрать тул?



# Популярные тулы для множественного выравнивания

- **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/>

# Как эти тулы запускать?

- Онлайн через **веб-интерфейс** для выполнения небольших задач:
  - <https://www.ebi.ac.uk/Tools/msa/>
- **Отдельные программы** для больших задач:
  - JalView: <https://www.jalview.org/>
  - MEGA
- Из **bash терминала**: Command Line Interface (CLI), для больших задач и времени затрачиваемых задач
- Используя **языки программирования**, для полного контроля инпута/аутпута:
  - BioPython в Python
  - SciKit-Bio для простых выравниваний и парсинга в Python
  - msa пакет для R

# Задание 1. Множественное выравнивание

**Инструкция:**

[https://github.com/michtrofimov/hse data analysis MSA](https://github.com/michtrofimov/hse_data_analysis_MSA)

Кратко:

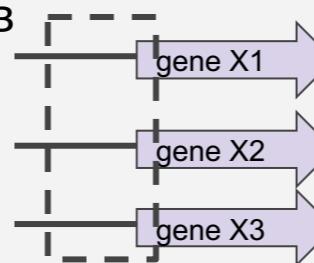
Скачайте файл с upstream регионами бактериальных ортологов `upstreams.fasta`.  
Выполните множественное выравнивание T-COFFEE, MUSCLE and CLUSTALW.  
Выберите наиболее консервативные участки без пропусков и сохраните `.fasta` file.

# Задача поиска мотивов: flowchart

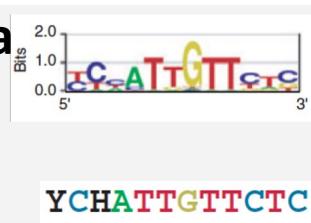
## выявление ко-регулируемых генов

- совместно экспрессирующиеся при определенных условиях и относящиеся к одной функциональной категории
- ортологичные гены
- ChIP-seq
- ...

## вытащить **upstream** регионы выявленных генов



## выбрать модель мотива (consensus или PWM) – чтобы получить мотивы и сравнить их



## Обнаружение мотивов

### «филогенетический след»:

выполнить MSA входных последовательностей  
(Множественное выравнивание последовательностей)

найти свободный от пробелов консервативный блок MSA

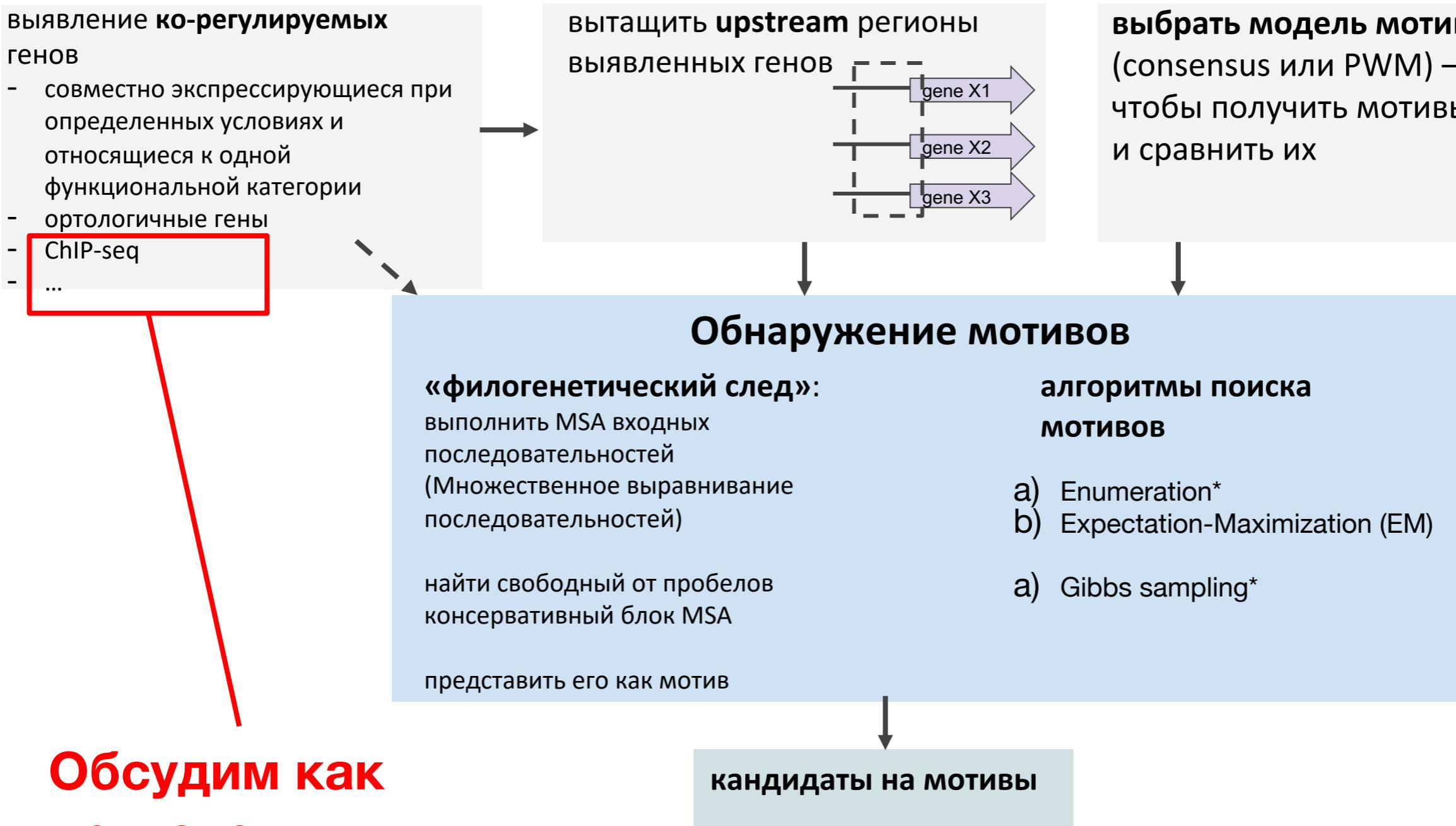
представить его как мотив

### алгоритмы поиска мотивов

- a) Enumeration\*
- b) Expectation-Maximization (EM)
- a) Gibbs sampling\*

кандидаты на мотивы

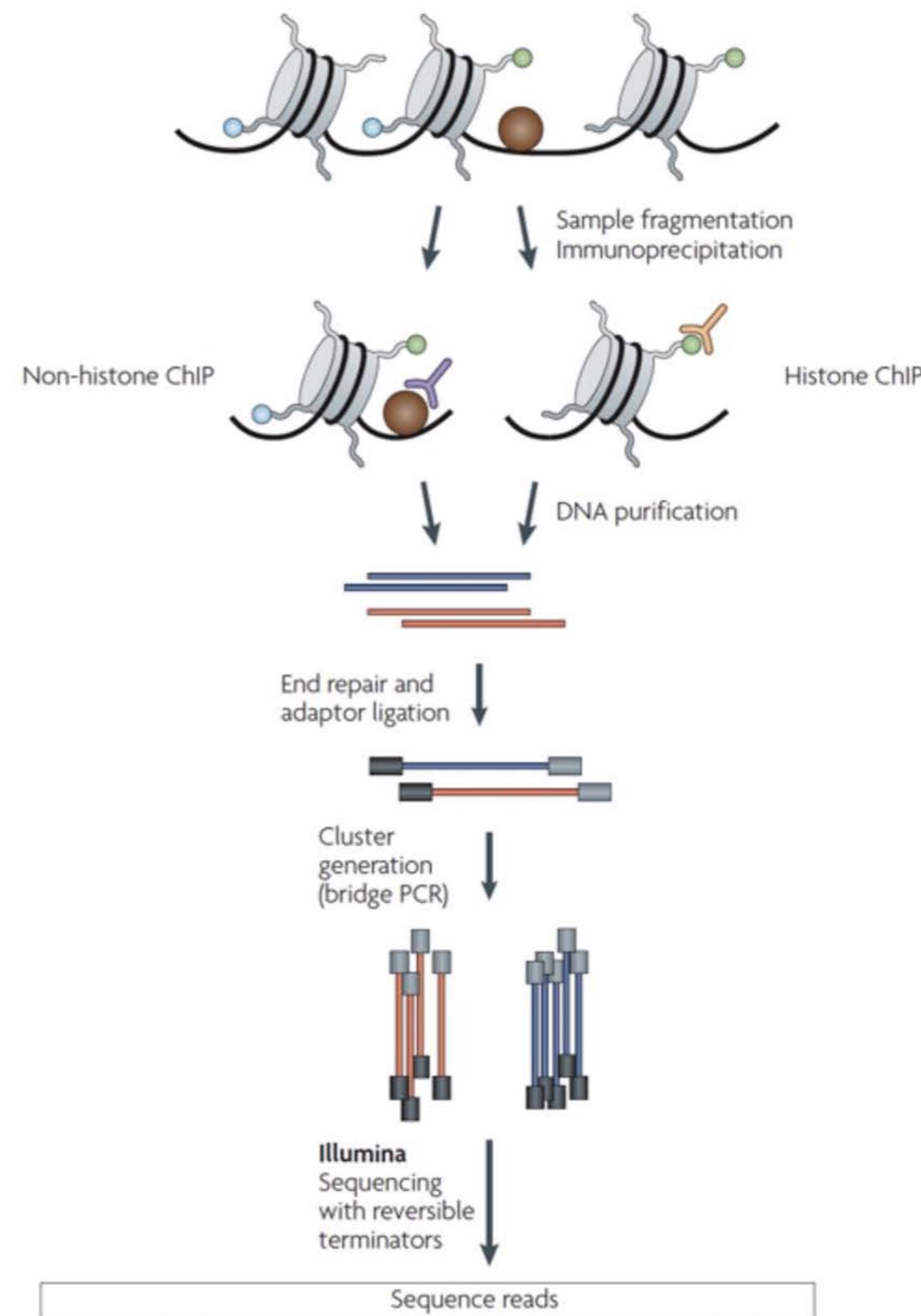
# Задача поиска мотивов: flowchart



**Обсудим как приготовить данные ChIP-seq для инпута в тул**

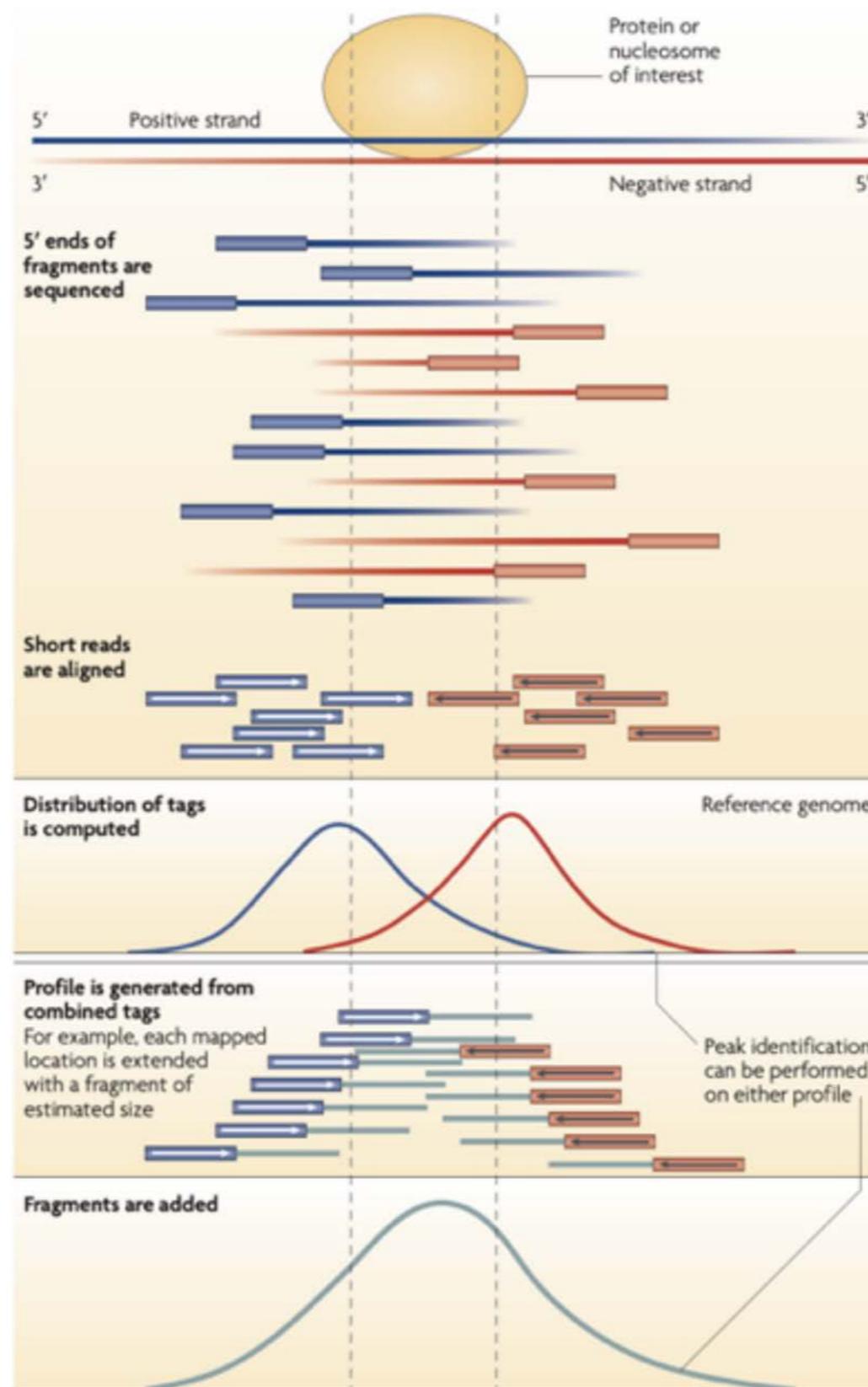
# База ChIP-seq: подготовка образцов

Chromatin-  
immunoprecipitation  
followed by sequencing:



# База ChIP-seq: анализ данных NGS

Binding events:



Read alignments:

Peak calling:

# Задача на поиск мотива

Имея множество последовательностей, найдите мотив (количество мотивов, их длина и локация во входных последовательностях)

# Задача на поиск мотива

- Попробуйте найти мотив

```
atgaccggatactgataaaaaaaaaagggggggggcgtacacattagataaacgtatgaagtacgttagactcgccgcgg  
acccctattttgagcagatttagtgacctggaaaaaaaaattgagtaaaaaactttccgaataaaaaaaaaaggggggga  
tgagtatccctggatgactaaaaaaaaagggggggtgctctccgattttgaatatgttaggatcattcgccagggtccga  
gctgagaattggatgaaaaaaaaagggggggtccacgcaatcgcaaccaacgcggacccaaaggcaagaccgataaaggaga  
tccctttcggtaatgtgccggaggctggtagtacgttaggaaagccctaacggacttaataaaaaaaaaaggggggcttatag  
gtcaatcatgttcttgtaatggattaaaaaaaaaggggggggaccgcttggcgcacccaaattcagtgtggcgagcgcaa  
cggtttggcccttggtagaggccccgtaaaaaaaaaggggggcaattatgagagagctaattatcgctgtgcgtttcat  
aacttgagttaaaaaaaaaggggggctgggcacataacaagaggagtcttcattatcagttatgttatgacactatgt  
ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcataaaaaaaaaaggggggaccgaaaggaaag  
ctggtagcaacgacagattttacgtgcattagctcgcttccgggatctaatacgacgaaagctttaaaaaaaaaggggggga
```

# Задача на поиск мотива

atgaccggatactgat **AAAAAAAAGGGGGGG** ggcgtacacattagataaacgtatgaagtacgttagactcggcgccg  
accctattttgagcagatttagtgcacctggaaaaaaaaattttagtacaactttccgaata **AAAAAAAAGGGGGGG** a  
tgagtatccctggatgactt **AAAAAAAAGGGGGGG** tgctctcccattttgaatatgttaggatcattcgccagggtccga  
gctgagaattggatg **AAAAAAAAGGGGGGG** tccacgcaatcgcaaccaacgcggacccaaaggcaagaccgataaaggaga  
tccctttcggtaatgtgccggaggctggttacgttaggaaagccctaacggacttaat **AAAAAAAAGGGGGGG** cttatag  
gtcaatcatgttcttgtaatggattt **AAAAAAAAGGGGGGG** gaccgcttggcgccccaaattcagtgtggcgagcgcaa  
cggtttggccctttagaggccccgt **AAAAAAAAGGGGGGG** caattatgagagagctaatctatcgctgtgttcat  
aacttgagtt **AAAAAAAAGGGGGGG** ctggggcacatacaagaggagtcttccttatcagttatgcttatgacactatgta  
ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatt **AAAAAAAAGGGGGGG** accgaaaggaaag  
ctggtagcaacgacagattttacgtcattagctcgcttccgggatctaatacgacgacgtt **AAAAAAAAGGGGGGG** a

# Задача на поиск мотива

- А если мы добавим SNPы:

atgaccggatactgat**AgAAGAAAGGttGGG**ggcgtacacattagataaacgtatgaagtacgttagactcgccgcgg  
accctat~~ttttt~~gagcagatttagtgacctggaaaaaaaaattgagtaaaaactttccgaata**CAAtAAAAcGGcGGG**a  
tgagtatccctggatgactt**AAAAtAAtGGaGtGG**tgctctcccgat~~ttt~~gaatatgttaggatcattcgccagggtccga  
gctgagaattggatg**CAAAAAAAGGGattG**tccacgcaatcgcaaccaacgcggacccaaaggcaagaccgataaaggaga  
tccctttgcgtaatgtgccggaggctggttacgttaggaagccctaacggacttaat**AtAAAtAAAGGaaGGG**cttata  
gtcaatcatgttcttgtaatggattt**AACAAAtAAGGGctGG**gaccgcttggcgccccaaattcagtgtggcgagcgcaa  
cggtttggccctttagaggccccgt**AtAAACAGGAAGGGc**caattatgagagagctaattatcgctgcgtttcat  
aacttgagtt**AAAAAAAtAGGGaGcc**ctggggcacataacaagaggagttttatcagttatgctgtatgacactatgt  
ttggcccattggctaaaagccaacttgacaaatggaagatagaatccttgcatt**ActAAAAAAGGAGcGG**accgaaaggaaag  
ctggtagcaacgacagattttacgtgcattagctcgcttccggggatctaatacgacgaagctt**ActAAAAAAGGAGcGGG**a

# Задача на поиск мотива

- Всё легче когда **знаешь ответ**

```
atgaccggatactgatagaagaaaggttggggcgtacacattagataaacgttatgaagtacgttagactcggcgcccg  
acccctattttttagcagatttagtgacctggaaaaaaaaattttagtacaactttccgaatacaataaaaacggcgga  
ttagtatccctggatgactaaaataatggagtggtgctctccgattttgaatatgttaggatcattcgccagggtccga  
gctgagaattggatgcaaaaaaaggattgtccacsgcaatcgcaaccaacgcggacccaaaggcaagaccgataaaggaga  
tccctttgcgtaatgtgccggaggctggttacgttaggaaagccctaacggacttaatataataaaggaaggcttata  
gtcaatcatgttcttgtaatggatttaacaataaggctggaccgcttggcgcacccaaattcagtgtggcgagcgcaa  
cggtttggccctttagaggccccgtataaacaaggaggccaattatgagagagactatctatcgctgcgttcat  
aacttgagttaaaaatagggagccctgggcacatacaagaggagtcttccttatcagttaatgttatgacactatgt  
ttggcccattggctaaagcccaacttgacaaatggaagatagaatcctgcataactaaaaaggagcggaccgaaaggaaag  
ctggtagcaacgacagattttacgtgcattagctcgcttccgggatctaatacgacgaagcttactaaaaaggagcggaa
```

# Проблема при поиске мотивов

По набору последовательностей найти мотив (количество мотивов, ширина каждого мотива и его расположение во входных последовательностях)

Проблемы:

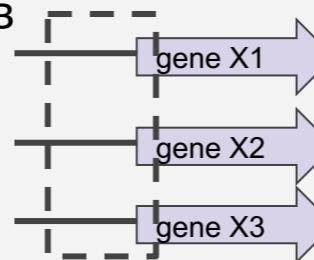
- входные последовательности могут быть длинными (до тысяч и миллионов)
- мотивы короткие и могут быть лишь слегка похожими (из-за замен)
- необходимо отличить мотив («сигнал») от геномного шума (неинформационной фоновой ДНК)

# Задача поиска мотивов: flowchart

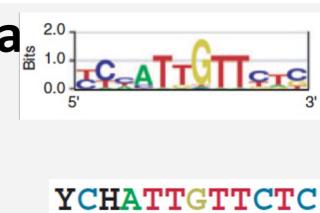
## выявление ко-регулируемых генов

- совместно экспрессирующиеся при определенных условиях и относящиеся к одной функциональной категории
- ортологичные гены
- ChIP-seq
- ...

## вытащить **upstream** регионы выявленных генов



**выбрать модель мотива** (consensus или PWM) – чтобы получить мотивы и сравнить их



## Обнаружение мотивов

### «филогенетический след»:

выполнить MSA входных последовательностей  
(Множественное выравнивание последовательностей)

найти свободный от пробелов консервативный блок MSA

представить его как мотив

### алгоритмы поиска мотивов

- a) Enumeration\*
- b) Expectation-Maximization (EM)
- a) Gibbs sampling\*

## кандидаты на мотивы

**Чтобы сравнивать, оценивать, ранжировать мотивы, нам нужна метрика оценки и модель (способ представления) для мотивов**

# Модели мотивов

ТАТААТ  
ТААААТ  
ТААТАТ  
ТГТААТ  
ТАТАСТ

– набор кандидатов на мотив

- консенсусная последовательность –

Т[АГ][АТ][АТ][АС]Т

- позиционная матрица частот (PFM), or позиционная матрица каунтов (PCM) –

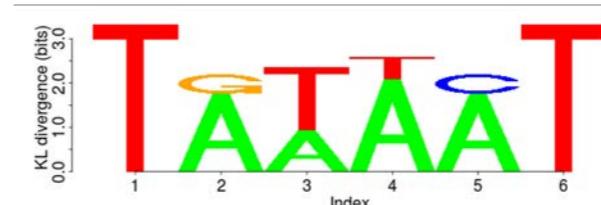
	1	2	3	4	5	6
A	0	4	2	4	4	0
C	0	0	0	0	1	0
G	0	1	0	0	0	0
T	5	0	3	1	0	5

- позиционная матрица вероятностей (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

- позиционно-специфичная матрица весов (PWM)

- матрица контента или лого последовательности –



# Модели мотивов: консенсус

В консенсусной последовательности перечислены нуклеотиды, которые разрешены в данной позиции. Рассмотрим следующий блок выравнивания без пробелов:

TATAAT

TAATAAT

TAATAT

TGTAAT

TATACT

**Консенсус:** T[AG][AT][AT][AC]T

**Проблемы:**

- Не позволяет учесть различные предпочтения для разных нуклеотидов
- Не позволяет учесть фоновые частоты нуклеотидов

# Модели мотивов

---

123456

ТАТААТ

ТААААТ

ТААААТ

TGTAAT

ТАТАСТ

Полученное множество  
кандидатов на мотив из  
множественного  
выравнивания

# Модели мотивов: позиционная матрица каунтов (частот)

123456	position <b>count</b> matrix PCM (position <b>frequency</b> matrix PFM)						
TATAAT	1	2	3	4	5	6	
TAAAAT	A	0	4	2	4	4	0
TAATAT	C	0	0	0	1	0	
TGTAAT	G	0	1	0	0	0	0
TATACT	T	5	0	3	1	0	5



😊 PCM (PFM) подсчитывает количество вхождений каждого нуклеотида в каждую позицию. Мы лучше понимаем различные предпочтения для разных нуклеотидов

😢 PCM (PFM) зависит от количества последовательностей, которые были первоначально выровнены

# Модели мотивов: позиционная матрица вероятностей

	position <b>count</b> matrix PCM (position <b>frequency</b> matrix PFM)						position <b>probability</b> matrix PPM					
123456	1	2	3	4	5	6	1	2	3	4	5	6
TATAAT	A	0	4	2	4	4	0	0.0	0.8	0.4	0.8	0.8
TAAAAT	C	0	0	0	0	1	0	0.0	0.0	0.0	0.0	0.2
TAATAT	G	0	1	0	0	0	0	0.0	0.2	0.0	0.0	0.0
TGTAAT	T	5	0	3	1	0	5	1.0	0.0	0.6	0.2	0.0
TATACT												1.0

count / (number of input sequences)

😊 PPM нормализует матрицу каунтов **по количеству наблюдений**, в результате чего получается оценка вероятности наблюдения каждой буквы в данной позиции  
⇒ Представление мотивов больше **не зависит от количества выровненных последовательностей**

# Модели мотивов: позиционная матрица вероятностей

	position <b>count</b> matrix PCM (position <b>frequency</b> matrix PFM)						position <b>probability</b> matrix PPM					
123456	1	2	3	4	5	6	1	2	3	4	5	6
TATAAT	A	0	4	2	4	4	0	0.0	0.8	0.4	0.8	0.8
TAAAAT	C	0	0	0	0	1	0	0.0	0.0	0.0	0.0	0.2
TAATAT	G	0	1	0	0	0	0	0.0	0.2	0.0	0.0	0.0
TGTAAT	T	5	0	3	1	0	5	1.0	0.0	0.6	0.2	0.0
TATACT												1.0

count / (number of input sequences)

😊 PPM нормализует матрицу каунтов **по количеству наблюдений**, в результате чего получается оценка вероятности наблюдения каждой буквы в данной позиции ⇒ Представление мотивов больше **не зависит от количества выровненных последовательностей**

😢 он не дает нам представления о том, насколько «удивительным» было бы наблюдение любой данной последовательности, соответствующей мотиву. Нам нужно оценить вероятность того, что этот наблюдаемый паттерн может быть случайно найден в геноме. Нам нужно отличить информативный паттерн (например, специфическое связывание, распознаваемое TF) от «неинформационного» геномного «шума» (неспецифические сайты, например, не распознаваемые TF). Мы должны учитывать этот шум

# Модели мотивов: позиционная матрица весов

	position <b>count</b> matrix PCM (position <b>frequency</b> matrix PFM)						position <b>probability</b> matrix PPM						
123456	1	2	3	4	5	6	1	2	3	4	5	6	
TATAAT	1	2	3	4	5	6	1	2	3	4	5	6	
TAAAAT	A	0	4	2	4	4	0	0.0	0.8	0.4	0.8	0.8	0.0
TAATAT	C	0	0	0	0	1	0	0.0	0.0	0.0	0.0	0.2	0.0
TGTAAT	G	0	1	0	0	0	0	0.0	0.2	0.0	0.0	0.0	0.0
TATACT	T	5	0	3	1	0	5	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

	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

Под «бэкграундом» здесь понимается состав оснований в неспецифических сайтах (т.е. в последовательностях, которые не обязательно связывают ЦГ). Фон является однородным ( $b_n = 1/4$ ) или геномными частотами

# Модели мотивов: позиционная матрица весов

	position <b>count</b> matrix PCM (position <b>frequency</b> matrix PFM)						position <b>probability</b> matrix PPM						
123456	1	2	3	4	5	6	1	2	3	4	5	6	
TATAAT	1	2	3	4	5	6	1	2	3	4	5	6	
TAAAAT	A	0	4	2	4	4	0	0.0	0.8	0.4	0.8	0.8	0.0
TAATAT	C	0	0	0	0	1	0	0.0	0.0	0.0	0.0	0.2	0.0
TGTAAT	G	0	1	0	0	0	0	0.0	0.2	0.0	0.0	0.0	0.0
TATACT	T	5	0	3	1	0	5	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

	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

Под «бэкграундом» здесь понимается состав оснований в неспецифических сайтах (т.е. в последовательностях, которые не обязательно связывают ЦГ). Фон является однородным ( $b_n = 1/4$ ) или геномными частотами

# Модели мотивов: позиционная матрица весов

	position <b>count</b> matrix PCM (position <b>frequency</b> matrix PFM)						position <b>probability</b> matrix PPM						
123456	1	2	3	4	5	6	1	2	3	4	5	6	
TATAAT	1	2	3	4	5	6	1	2	3	4	5	6	
TAAAAT	A	0	4	2	4	4	0	0.0	0.8	0.4	0.8	0.8	0.0
TAATAT	C	0	0	0	0	1	0	0.0	0.0	0.0	0.0	0.2	0.0
TGTAAT	G	0	1	0	0	0	0	0.0	0.2	0.0	0.0	0.0	0.0
TATACT	T	5	0	3	1	0	5	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

	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

Под «бэкграундом» здесь понимается состав оснований в неспецифических сайтах (т.е. в последовательностях, которые не обязательно связывают ЦГ). Фон является однородным ( $b_n = 1/4$ ) или геномными частотами

❗ Но у нас бесконечность в матрице!

В небольших наборах данных всегда есть шанс, что возможное событие не произойдет (нули в матрице подсчетов/частот -> бесконечность в матрице весов).

Чтобы учесть редкие события и устранить эмпирические нулевые частоты, мы используем псевдоколаунты. Добавим псевдоколоунт к каждому счету в матрице счетов

# Модели мотивов: позиционная матрица весов

	position <b>count</b> matrix PCM (position <b>frequency</b> matrix PFM)						position <b>probability</b> matrix PPM					
123456	1	2	3	4	5	6	1	2	3	4	5	6
TATAAT	1	2	3	4	5	6	1	2	3	4	5	6
TAAAAT	A	0	4	2	4	4	0	0.0	0.8	0.4	0.8	0.8
TAATAT	C	0	0	0	0	1	0	0.0	0.0	0.0	0.0	0.2
TGTAAT	G	0	1	0	0	0	0	0.0	0.2	0.0	0.0	0.0
TATACT	T	5	0	3	1	0	5	1.0	0.0	0.6	0.2	0.0

$$M_{p,n} = \log_2 \left( \frac{p_{p,n}}{b_n} \right)$$

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

$b_n$  is probability of nucleotide  $n$  in background

	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



position **weight** matrix PWM

# Модели мотивов: лого последовательности

Относительная энтропия (расстояние Куллбэка-Лейблера) сайта связывания по отношению к фоновым частотам:

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

the frequency of base b at position i  
the background frequency of base b in the genome

Относительная энтропия измеряет степень несогласия (несходства) между наблюдаемыми и фоновыми частотами оснований и, таким образом, может быть использована для расчета значимости самого мотива

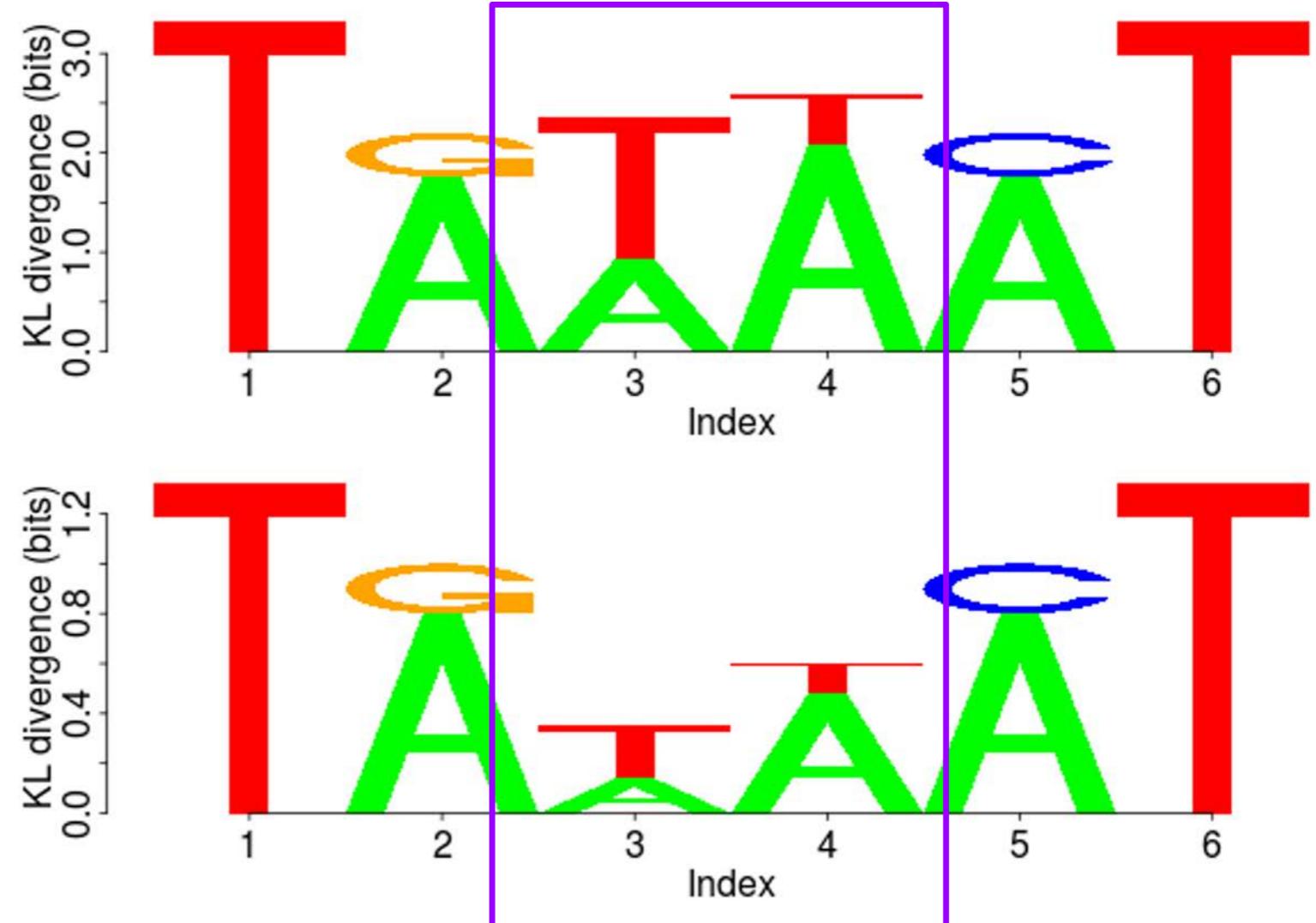
Высота каждого столбца - значимость позиции (несходство с бэкграундом).  
Относительный размер буквы - частота встречаемости нуклеотида.

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

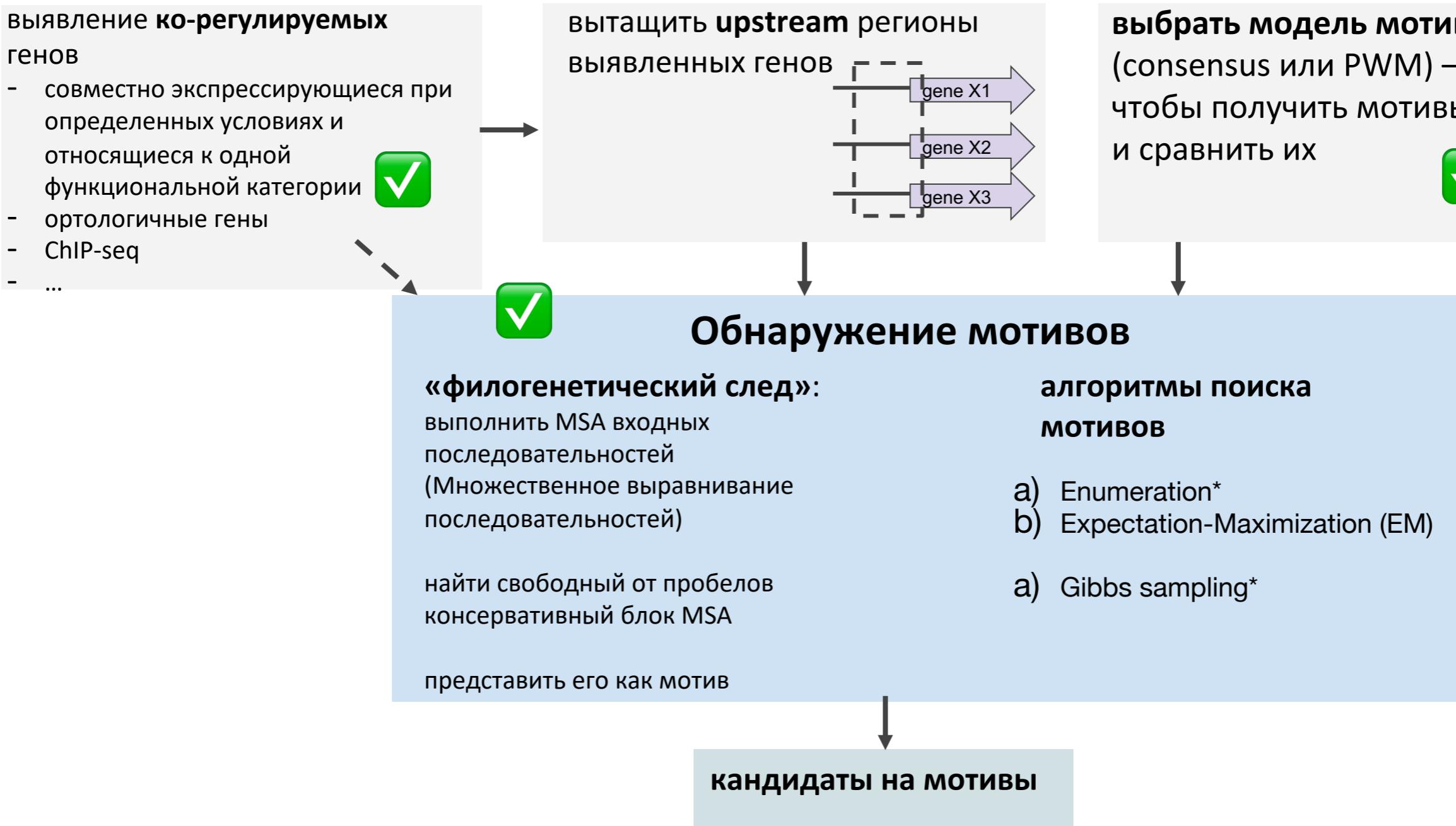
GC-rich  
background:

появление Т и А более значительно на бэкграунде, богатом GC, чем на фоне, богатом AT (=low-GC)

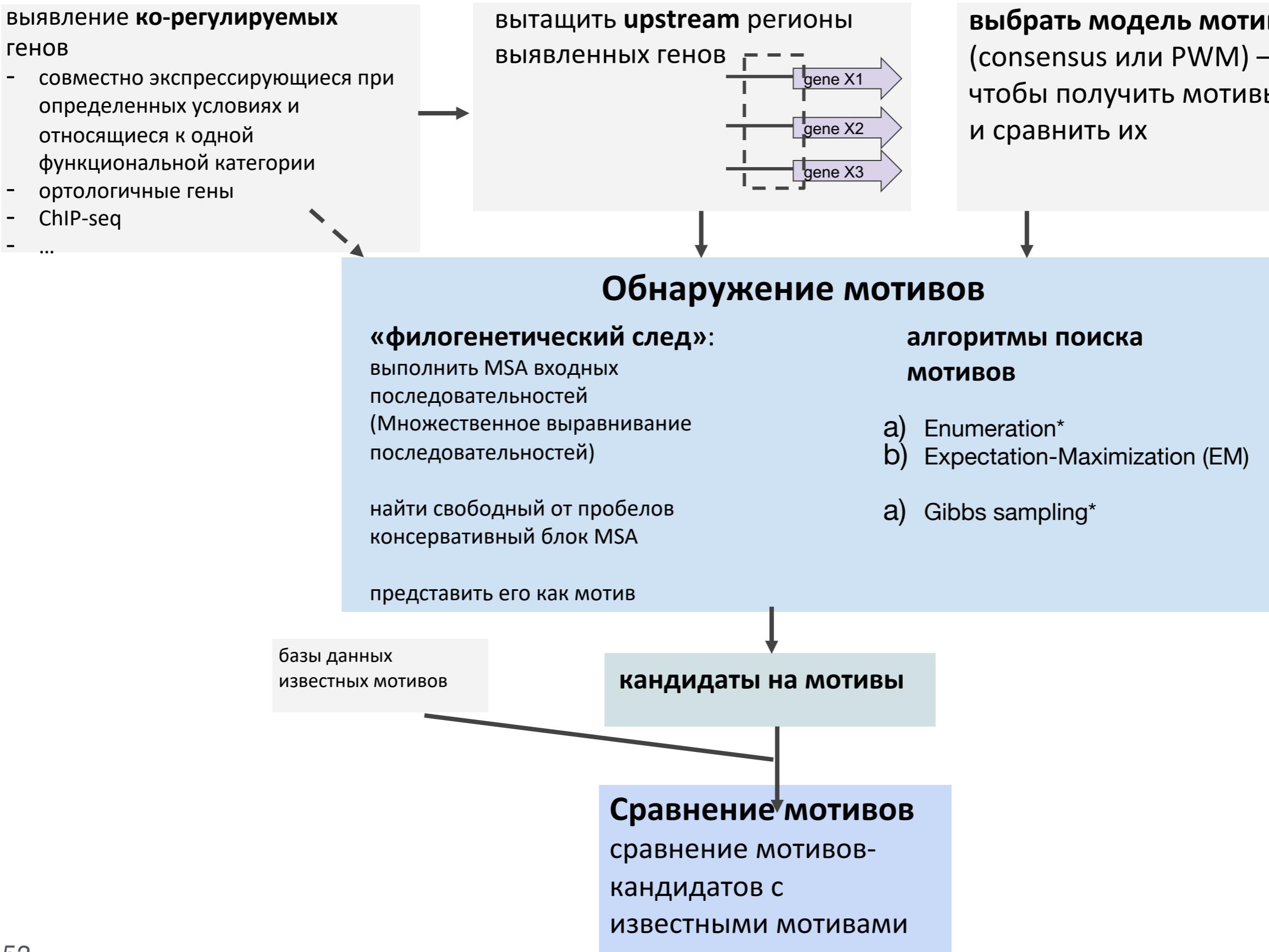
AT-rich  
background:



# Задача поиска мотивов: flowchart



# Задача поиска мотивов: flowchart

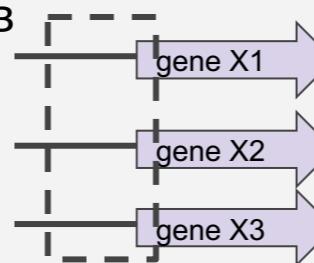


# Задача поиска мотивов: flowchart

## выявление ко-регулируемых генов

- совместно экспрессирующиеся при определенных условиях и относящиеся к одной функциональной категории
- ортологичные гены
- ChIP-seq
- ...

вытащить **upstream** регионы выявленных генов



выбрать модель мотива (consensus или PWM) – чтобы получить мотивы и сравнить их



УСНАТТГТТСТС

## Обнаружение мотивов

### «филогенетический след»:

выполнить MSA входных последовательностей  
(Множественное выравнивание последовательностей)

найти свободный от пробелов консервативный блок MSA

представить его как мотив

### алгоритмы поиска мотивов

- a) Enumeration\*
- b) Expectation-Maximization (EM)
- c) Gibbs sampling\*

базы данных известных мотивов

### кандидаты на мотивы

Сравнение мотивов  
сравнение мотивов-кандидатов с известными мотивами

**Задания 2-4**

# Тулы

## Онлайн тулы:

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

## CLI тулы:

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

## Тул в питоне:

- BioPython motifs

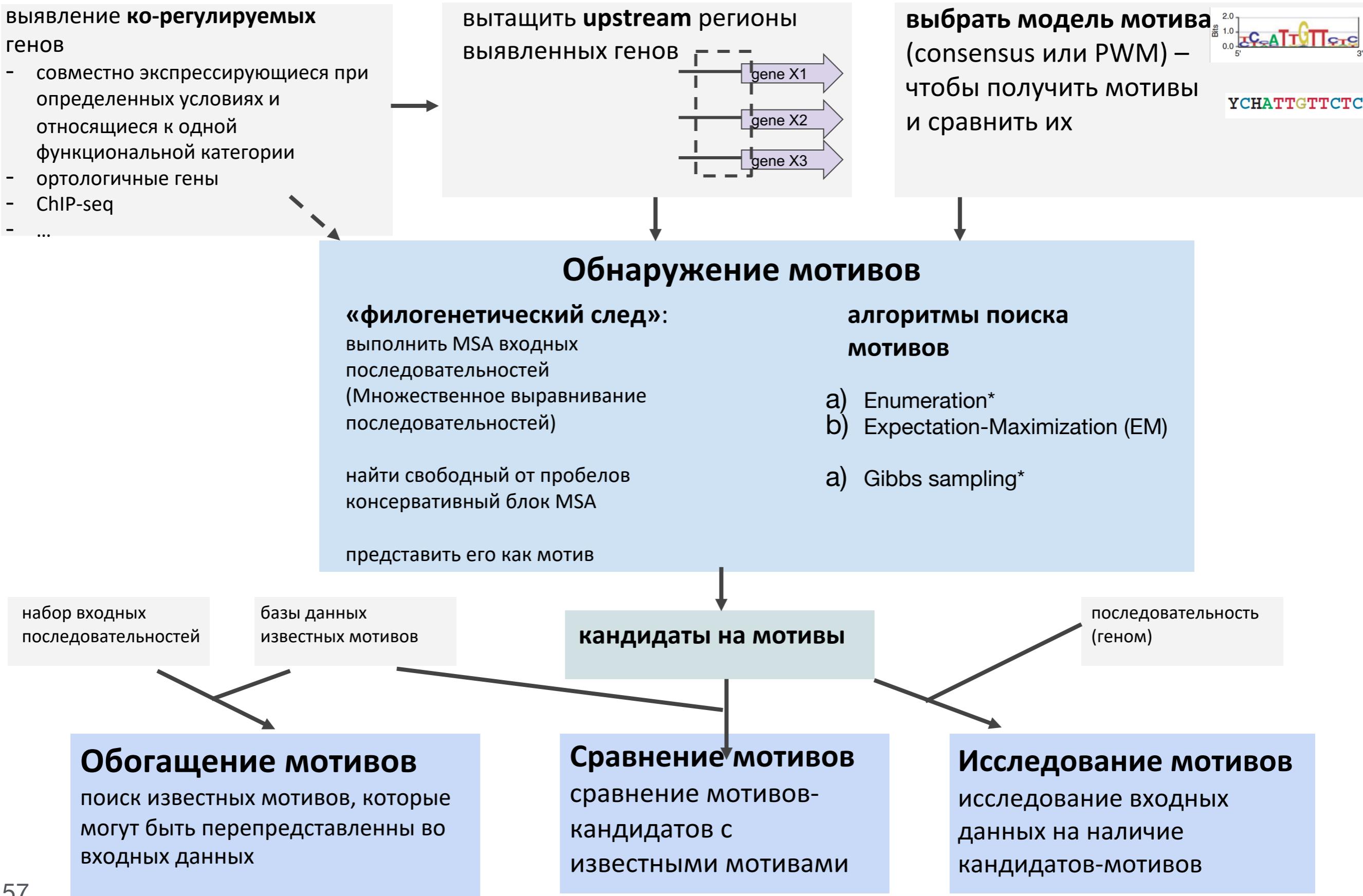
# Задания 2-4. Поиск мотивов

**Инструкция:** [https://github.com/michtrofimov/hse\\_data\\_analysis\\_MSA](https://github.com/michtrofimov/hse_data_analysis_MSA)

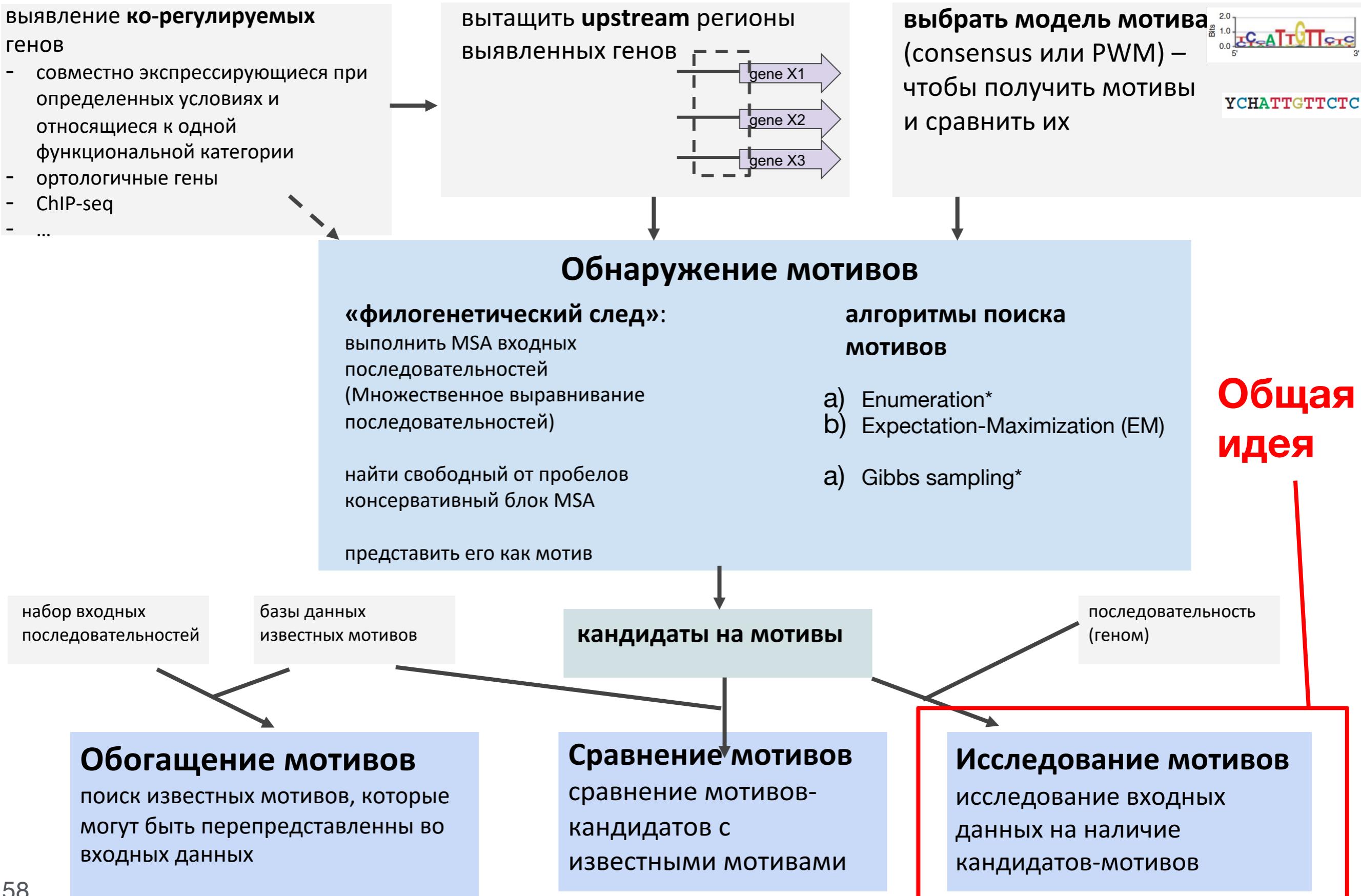
## Кратко

- Создайте матрицы каунтов, частот, весов и лого последовательности из выравнивания без пробелов с помощью инструментов RSAT:  
<http://embnet.ccg.unam.mx/rsat/> -> Matrix tools.
- Обработайте тот же набор последовательностей upstreams.fasta с помощью MEME: <http://meme-suite.org/>. Установите возможную длину мотива от 5 до 15. Похож ли результат на тот, что вы получили вручную?
- Загрузите файл с последовательностями пиков из ChIP-Seqа курицы (peaks.fasta). Найдите мотивы с помощью MEME-ChIP (<http://meme-suite.org/> -> MEME-ChIP). Какой белок использовался для ChIP-Seq?
- Повторите для вашего файла пиков, назначенного вам

# Задача поиска мотивов: flowchart



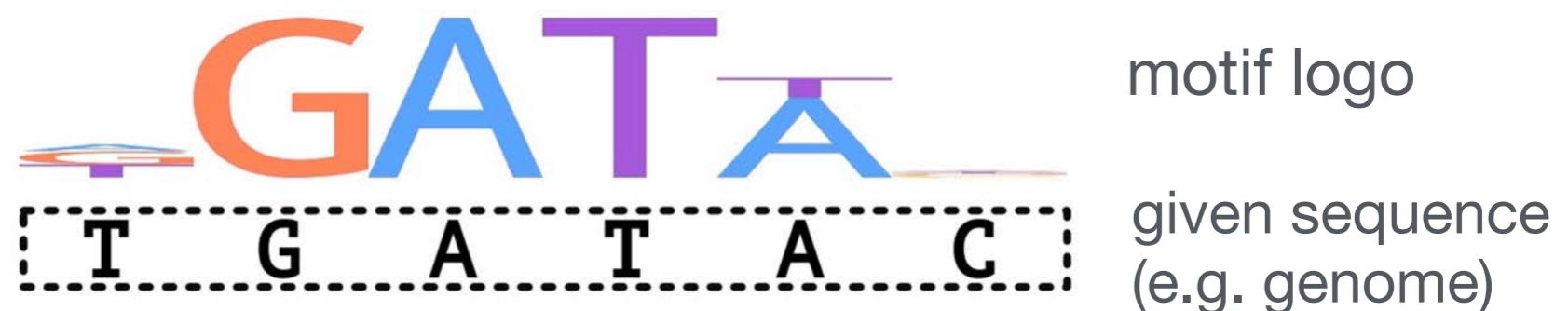
# Задача поиска мотивов: flowchart



# Сканирование мотивов

- Представим, что нам известен конкретный мотив и его PWM для некоторого белка. Как найти сайты связывания этого белка в геноме?

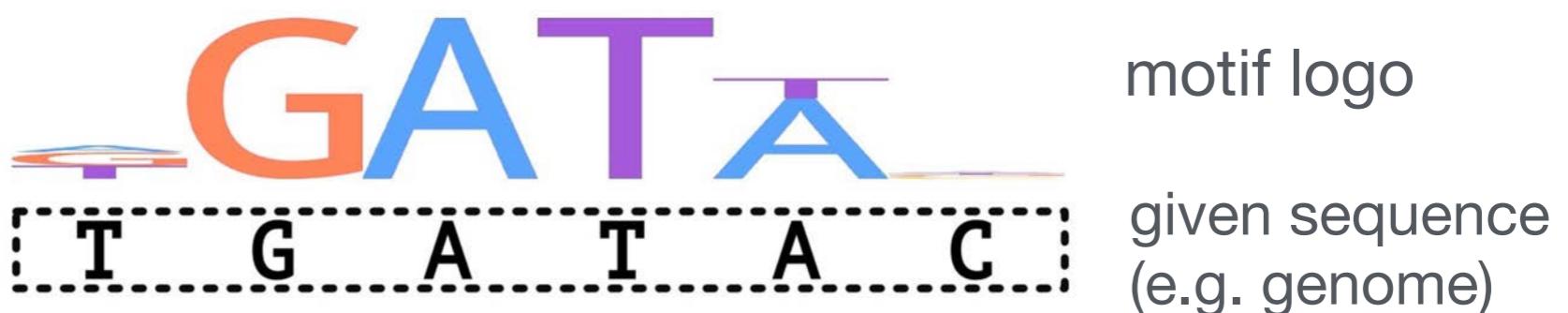
input data:



# Сканирование мотивов

- Представим, что нам известен конкретный мотив и его PWM для некоторого белка. Как найти сайты связывания этого белка в геноме?

input data:



Для каждой позиции  
PWM выделите  
соответствующую  
букву

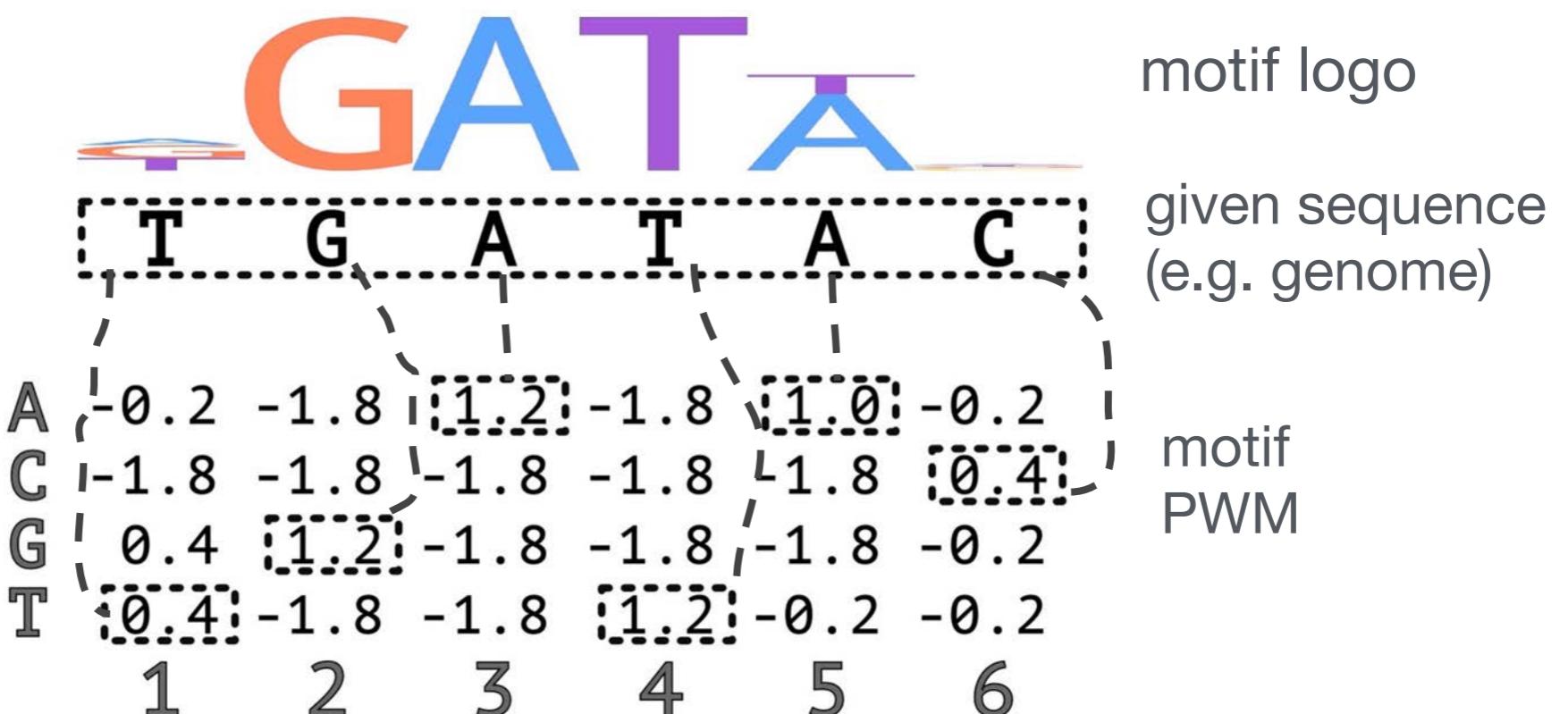
A	-0.2	-1.8	1.2	-1.8	1.0	-0.2
C	-1.8	-1.8	-1.8	-1.8	-1.8	0.4
G	0.4	1.2	-1.8	-1.8	-1.8	-0.2
T	0.4	-1.8	-1.8	1.2	-0.2	-0.2
	1	2	3	4	5	6

motif  
PWM

# Сканирование мотивов

- Представим, что нам известен конкретный мотив и его PWM для некоторого белка. Как найти сайты связывания этого белка в геноме?

input data:

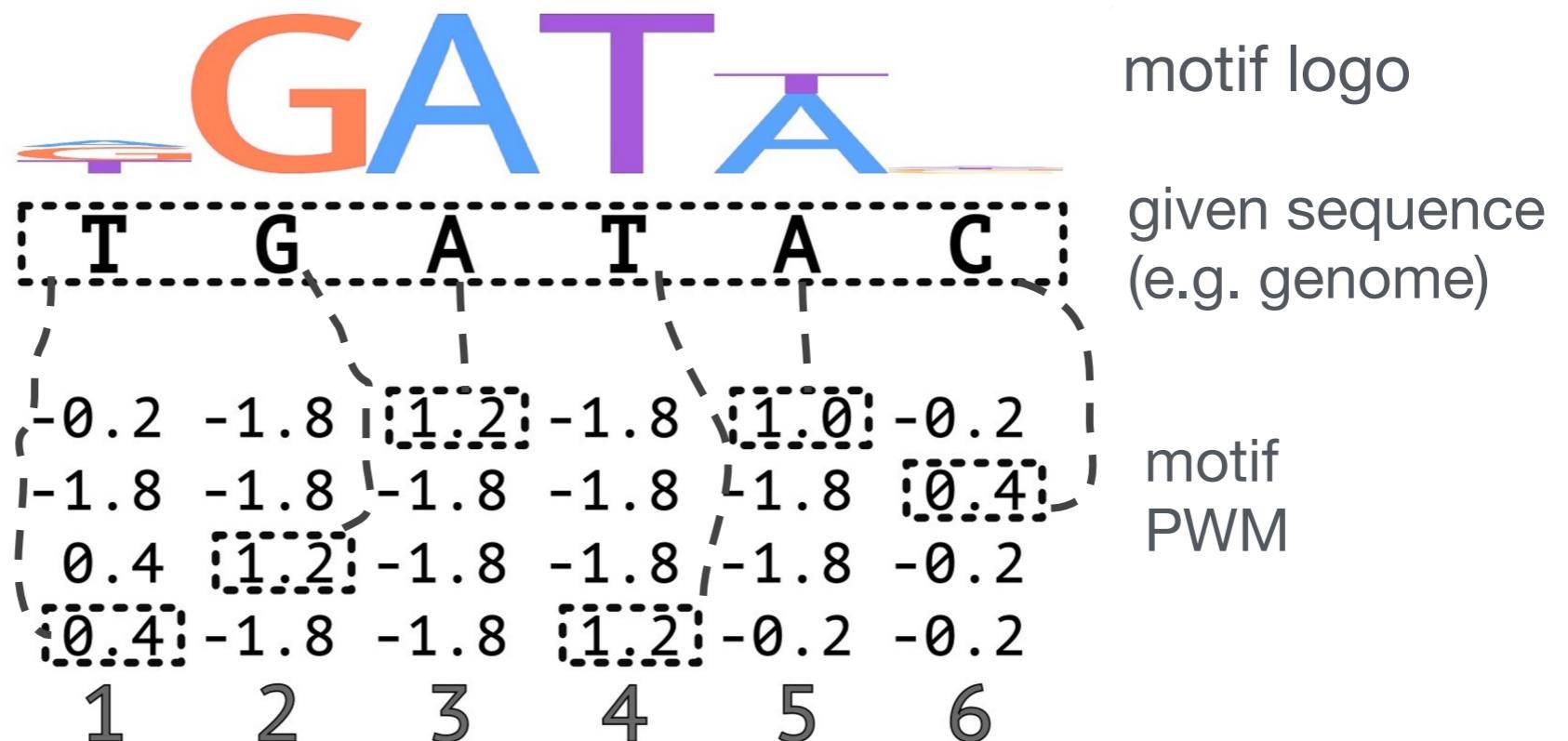


Для каждой позиции  
PWM выделите  
соответствующую  
букву

# Сканирование мотивов

- Представим, что нам известен конкретный мотив и его PWM для некоторого белка. Как найти сайты связывания этого белка в геноме?

input data:



Для каждой позиции  
PWM выделите  
соответствующую  
букву

Оценка  
последовательности  
для данного мотива  
показывает,  
насколько хорошо  
последовательность  
соответствует  
мотиву.

$$\sum [0.4 + 1.2 + 1.2 + 1.2 + 1.0 + 0.4] \quad \text{score}$$

$$Score(tGATAc) = 5.4$$

motif logo

given sequence  
(e.g. genome)

motif  
PWM

score

# Сканирование мотивов: выбор порога скора

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	PWM
C	-1.6	-1.6	0.00	-1.6	-1.6	-1.6	$S_{GGATTA} = 1.22 + 1.22 + 0.96 + 1.22 + 0.96 = \mathbf{6.8}$
G	1.22	1.22	-1.6	-1.6	-1.6	-1.6	$S_{GGGGGG} = 2.44 - 6.4 = \mathbf{-3.96}$
T	-1.6	-1.6	-1.6	1.22	1.22	0.00	$S = \mathbf{-9.6}$ the worst score

a)  $S_{min}$  в качестве порога:

False positive:  $S_{GGGGGG} = -3.96 > S_{min} = -9.6 \Rightarrow S_{GGGGGG}$  прошел, но не является истинным мотивом! не круто

a)  $S_{max}$  как порог:

$S_{GGGGGG} = -3.96 < S_{max} = 6.8 \Rightarrow S_{GGGGGG}$  отклонен и не является истинным мотивом, все в порядке

# Сканирование мотивов: выбор порога скора

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

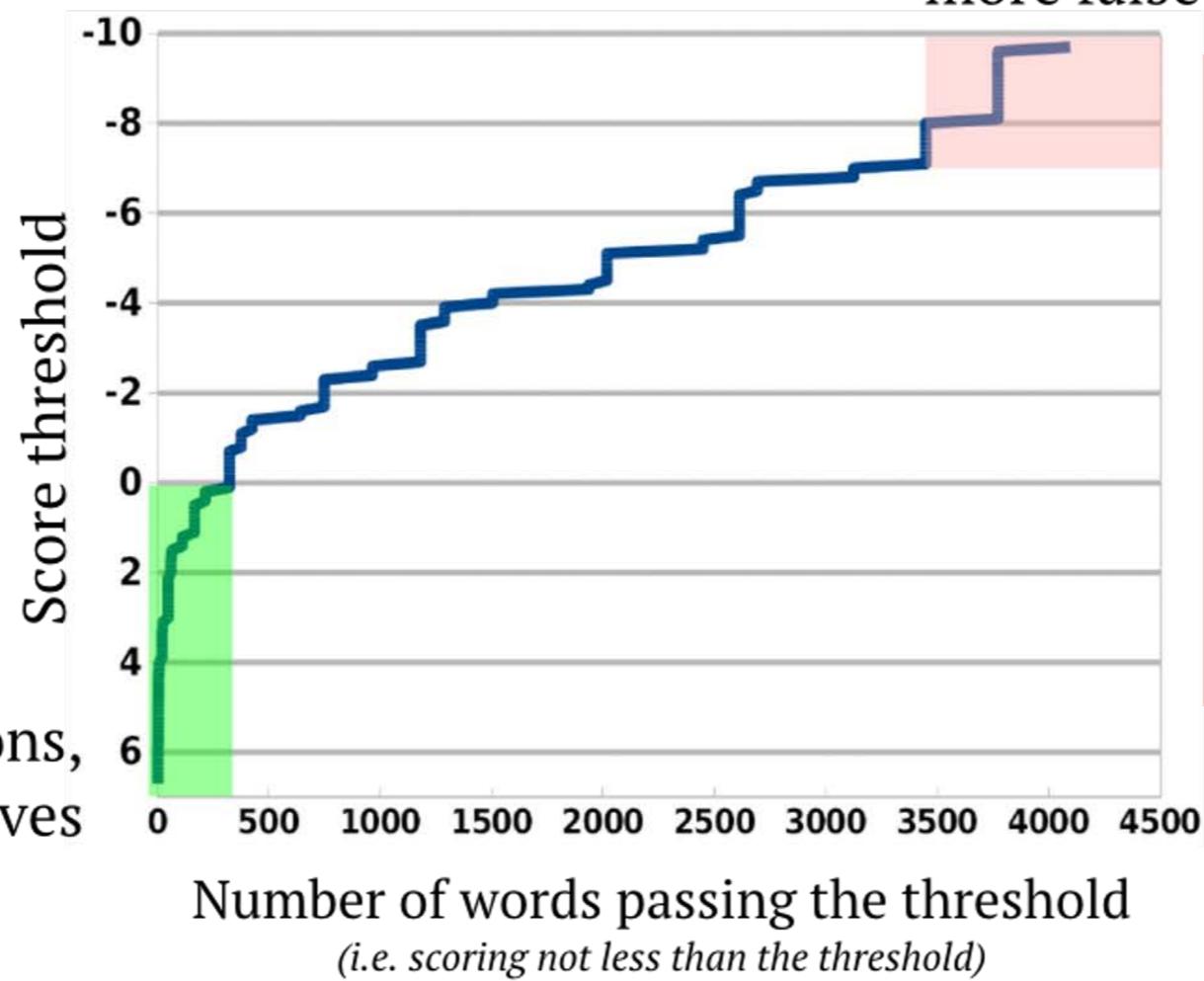
$$S_{\text{GGATTA}} = 1.22 + 1.22 + 0.96 + 1.22 + 0.96 = \mathbf{6.8}$$
$$S_{\text{GGGGGG}} = 2.44 - 6.4 = \mathbf{-3.96}$$

the best score

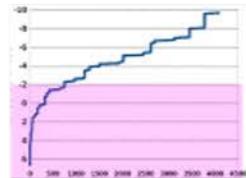
$S = -9.6$   
the worst score

порог слишком высок  
многие паттерны, которые являются настоящими мотивами, не проходят порог, и мы их теряем

less true positives

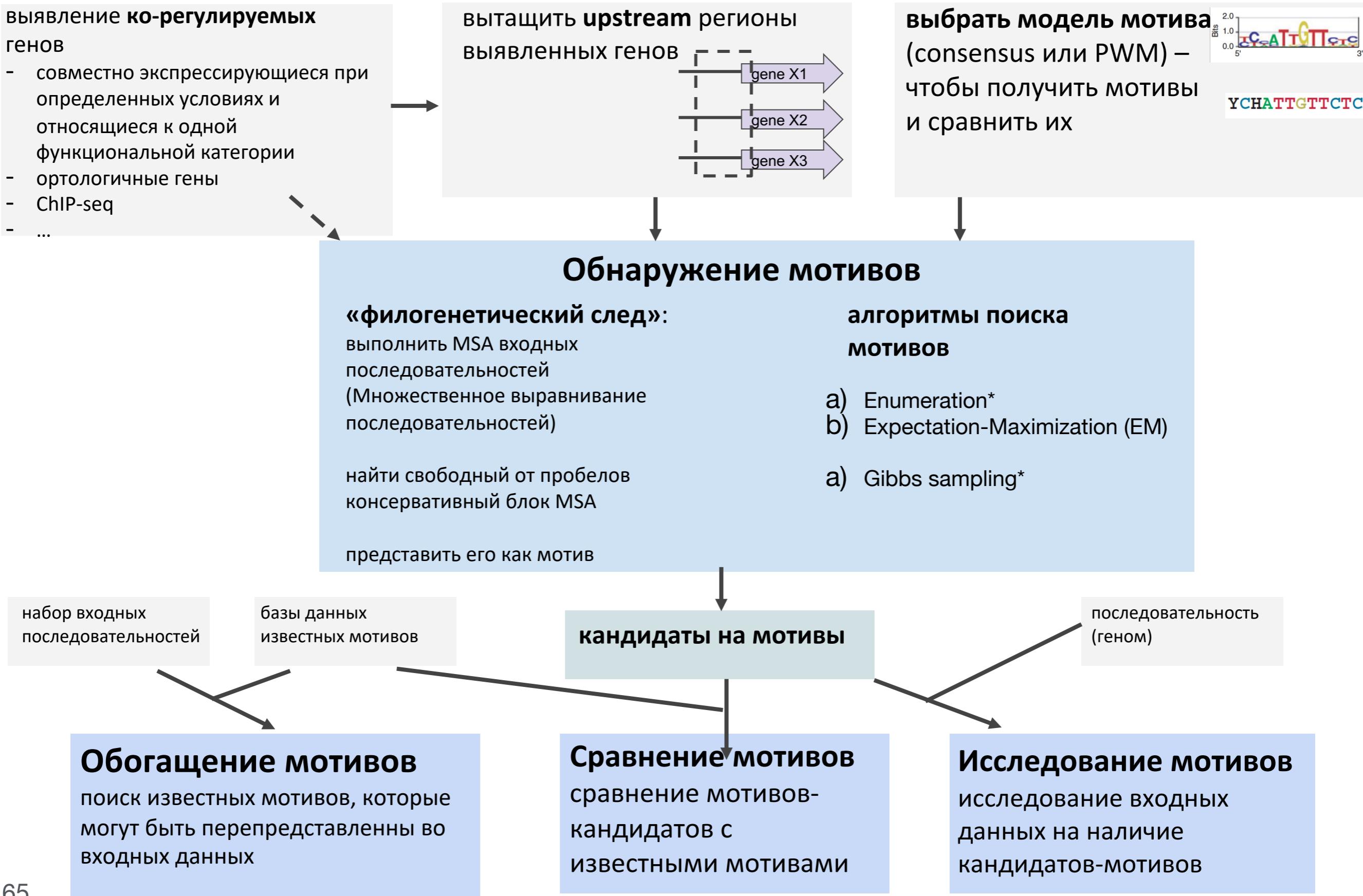


порог слишком низкий  
множество паттернов проходят порог и регистрируются как мотивы, но они не являются настоящими мотивами



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

# Задача поиска мотивов: flowchart

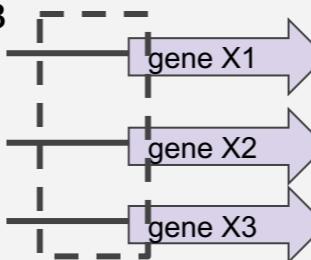


# Задача поиска мотивов: вычислительный подход

## выявление ко-регулируемых генов

- совместно экспрессирующиеся при определенных условиях и относящиеся к одной функциональной категории
- ортологичные гены
- ChIP-seq
- ...

вытащить **upstream** регионы выявленных генов



выбрать модель мотива (consensus или PWM) – чтобы получить мотивы и сравнить их



УСНАТТГТТСТС

## Обнаружение мотивов

### «филогенетический след»:

выполнить MSA входных последовательностей  
(Множественное выравнивание последовательностей)

### алгоритмы поиска мотивов

- a) Enumeration\*
- b) Expectation-Maximization (EM)
- a) Gibbs sampling\*



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

RSAT

найти свободный от пробелов консервативный блок MSA  
представить его как мотив

набор входных последовательностей

базы данных известных мотивов

## кандидаты на мотивы

последовательность (геном)

## Обогащение мотивов

поиск известных мотивов, которые могут быть перепредставлены во входных данных



сравнение мотивов-кандидатов с известными мотивами

## Исследование мотивов

исследование входных данных на наличие кандидатов-мотивов

# Домашнее задание

**Оформить пункты в единый .pdf документ следуя пунктам  
отсюда:**

**[https://github.com/michtrofimov/hse data analysis MSA](https://github.com/michtrofimov/hse_data_analysis_MSA)**

**В домашнем задании используйте понятный научный язык и не  
превышайте 200 слов на один ответ!**

# 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>