Сигналы и мотивы -3

Мотивы в последовательностях белков

I. Сигналы в ДНК. ПОВТОРЕНИЕ

- Что такое сигнал в ДНК
- Примеры сигналов в основных процессах в клетке
  - Репликация
  - Транскрипция
  - Трансляция
  - Регуляция транскрипции
- Методы поиска и описания сигналов
- Оценка представленности слова в геноме

Методы описания и поиска сигналов в ДНК

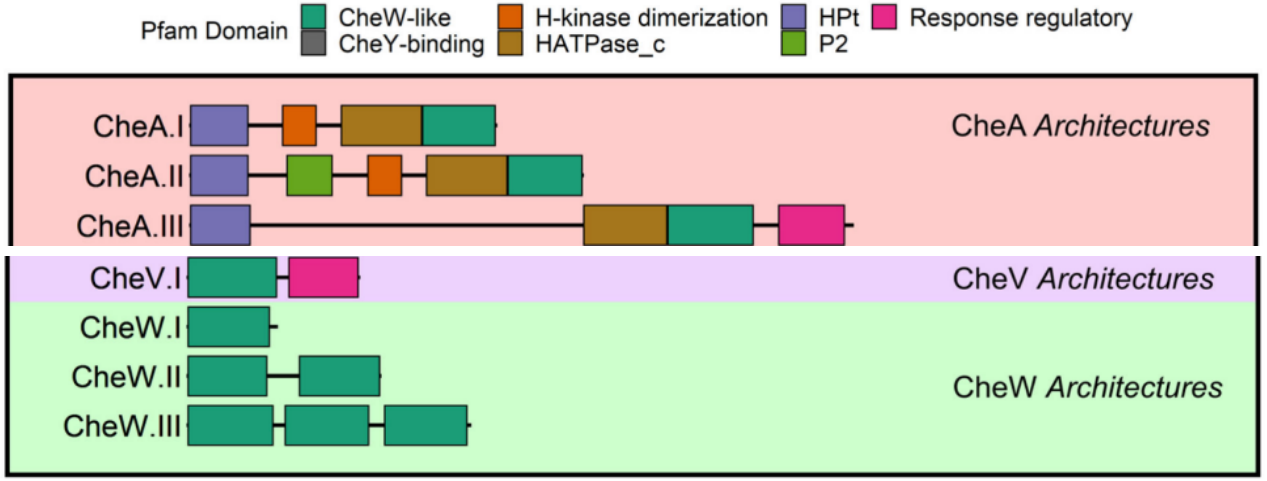
- PWM
  - Для какой задачи используется
  - Составление. Все этапы.
  - Поиск по PWM
  - Консенсус с учётом ambiguity codes
  - Паттерн
- Информационное содержание IC
  - Для чего используется
  - LOGO что такое, для чего нужно и как строится

- Поиск сигналов *de novo* (MEME и FIMO)
  - Как найти неизвестные сигналы в последовательностях
  - Как найти найденный сигнал во всех последовательностях

II Мотивы в доменах белков

Домен - единица непрерывной эволюции в белках

Доменные архитектуры белков, содержащих домен CheW-like
Из систем хемотаксиса прокариот



Мотив. Что такое? Короткие консервативные последовательности в гомологичных белках

```
   *       400         *       420         *       440         *       460
GSMPSGSPCTALLNSIINNVNLYYVFSKIFGKSPVFF.................CQALKILC.YGDDVLIVFSRDV
GGLPSGCAATSMLNTIMNNIIIRAGLYLTYKNFEFDD.....................VKVLS.YGDDLLVATNYQL
GGMPSGCSATSIINTILNNIYVLYALRRHYEGVELDT.....................YTMIS.YGDDIVVASDYDL
VGLSSGQGATDIMGTLLMSITYLVMQLDHTAPHLNSRIKDMPSACRFLDSYWQGHEEIRQIS.KSDDAILGWTKGR
RGNNSGQPSTVVDNSIMVVLAMHYALIKECVEFEEID.................STCVFFV.NGDDLLIAVNPEK
VGTQR.QPSTVVDNTLVLMTAFLYAYIHKTGDRELAL................LNERFIFVC.NGDDNKFAISPQF
GGGPSGFPMTVIFNSFINLFYLQSAWIMLARFNGRQDISH..........PCNFPKYVRACV.YGDDNIVAIKMEV
CGIPSGFPMTVIVNSIFNEILIRYHYKKLMREQCAPELMV..........QSFDKLIGLVT.YGDDNLISVNAVV
EGLPSGFPCTSQVNSINHWIITLCALSEATGLSPDVV.................QSMSYFSFYGDDEIVSTDIDF
RGLPSGMPFTSVINSICHWLLWSAAVYKSCAEIGLHCS.............NLYEDAPFYT.YGDDGVYAMTPMM
```

РНК зависимая РНК полимераза (RdRP)

Консервативные – значит важные

- активные центры ферментов
- участки, связывающие лиганды
- участки белок-белкового взаимодействия
- мотивы связывания с ДНК, РНК
- Мотивы связывания с другими молекулами и комплексами – например, вирионами
- И проч. …..

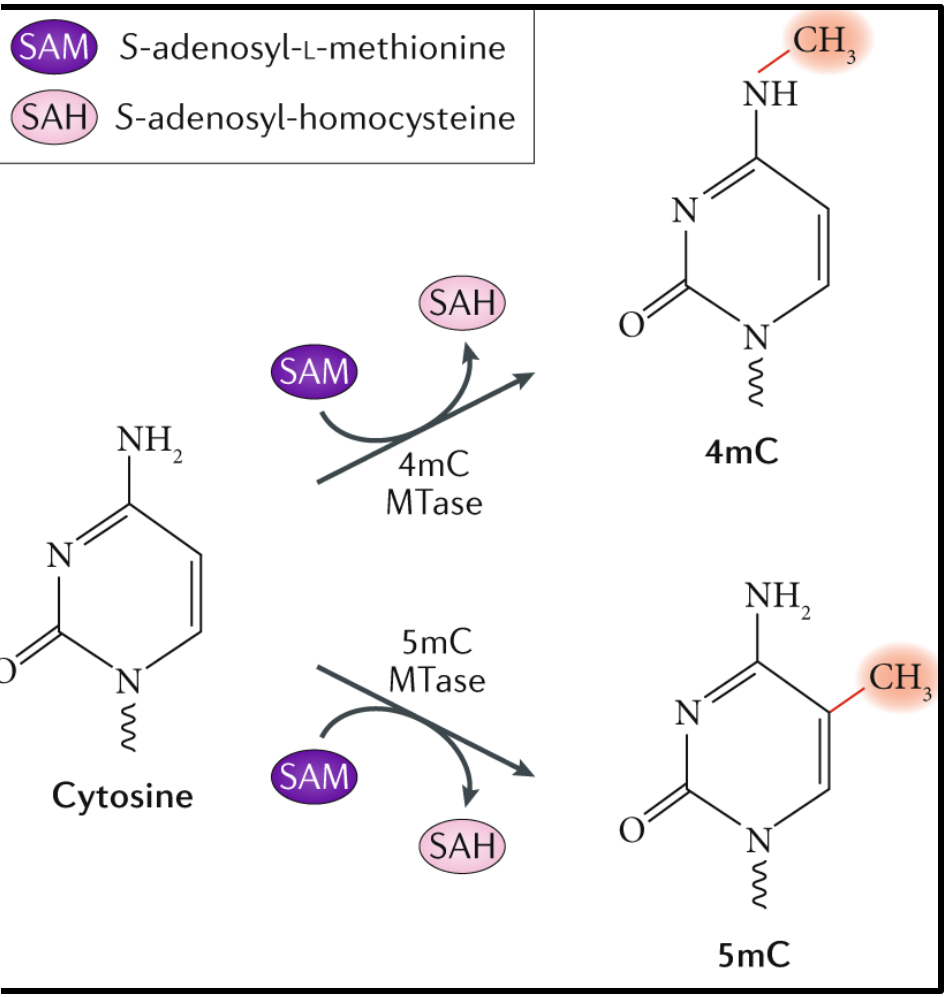
Поиск мотивов с помощью анализа множественного выравнивания

Используем Jalview

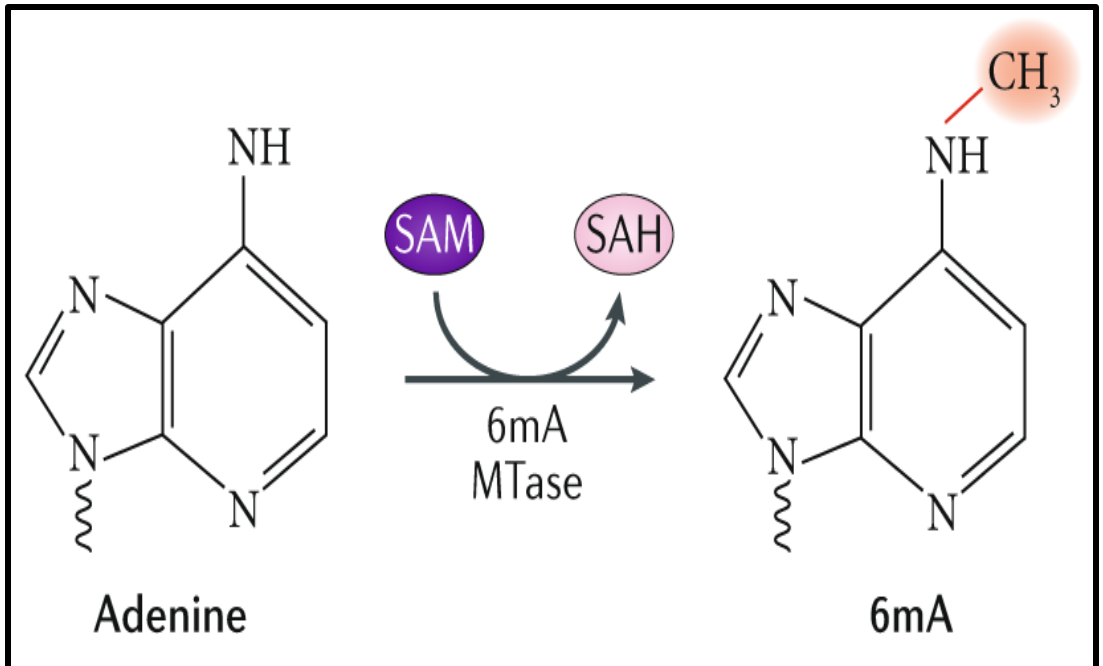
Пример.ДНК метилтрансферазы (МТазы)

1. Узнают короткую последовательность ДНК (например, GATC)
    1. Узнаваемая последовательность может различаться даже у гомологичных Мтаз
2. Метилируют определенное основание ДНК в сайте или рядом.
3. Три вида метилирования:
    1. 5mC метилируется углерод цитозина в положении 5
    2. 4mC метилируется азот цитозина в положении 4
    3. 6mA метилируется азот аденина в положении 6
4. Источником метильной группы служит SAM S-adenosile-L-methionine

Три вида метилирования ДНК



- 5mC
- 4mC
- 6mA



Каталитический домен МТазы

Ожидаем консервативность

- Активного центра у Мтаз с одним типом метилирования
- Участка связывания SAM

Демонстрация она же задание для выполнения в классе

Домен Pfam

C-5 cytosine-specific DNA methylase

PF00145

Задание

1. Выберите консервативный мотив в выравнивании доменов PF00145 из разных белков
2. Опишите его паттерном
3. Найдите все находки паттерна
    a. в выравнивании
    b. в банке SwissProt
4. Опишите результат в форме

ДЕМОНСТРАЦИЯ Pfam => Jalview

1. Скачайте выравнивание seed. Смените расширение на .msf
2. Откройте в Jalview

ДЕМОНСТРАЦИЯ В JALVIEW

2. Окраска Clustal. Подберите порог identity так чтобы видеть только самые консервативные мотивы.

3. Выберите на глаз один мотив с самым большим информационным содержанием

4. Составьте паттерн Jalview этого мотива.

5. Выполните поиск по этому паттерну во всем выравнивании. Число находок, из них – на месте

6. Заполните форму

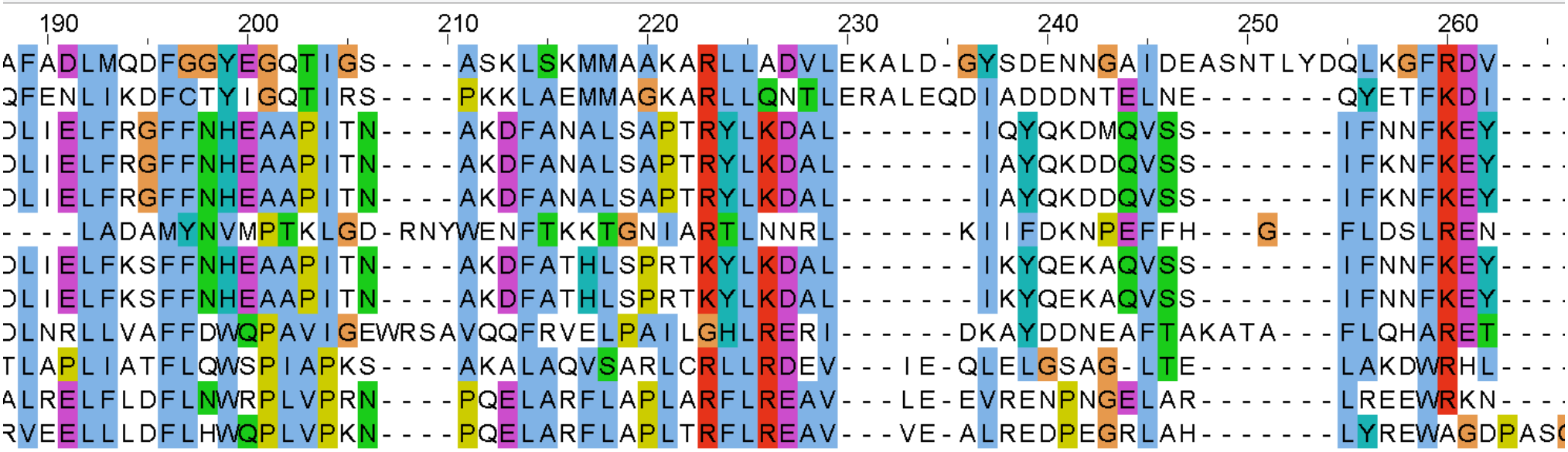
8. Опишите результат.

КОНЕЦ ЗАДАНИЯ

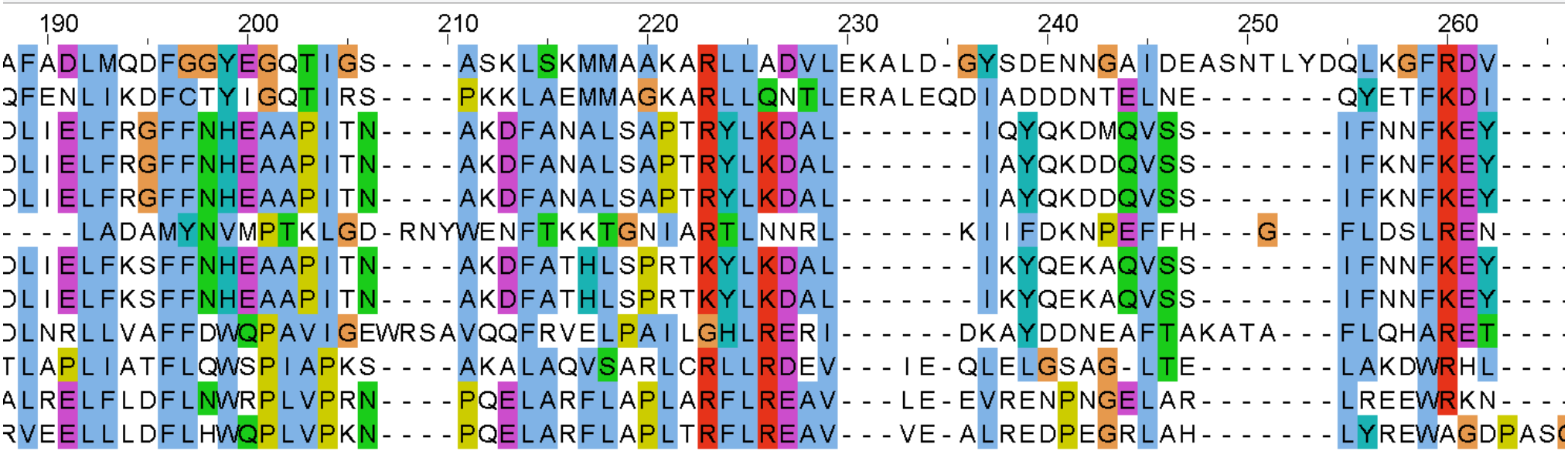
Слайды 14 -20 показывают то же, что в задании быстро просмотреть

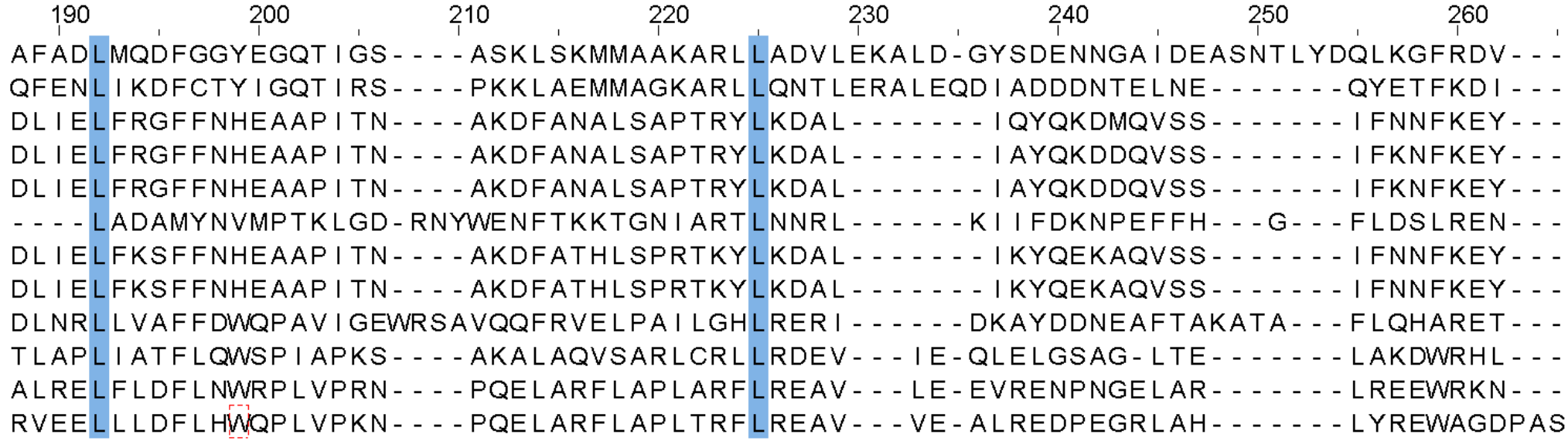
Выравнивание 12 Мтаз.

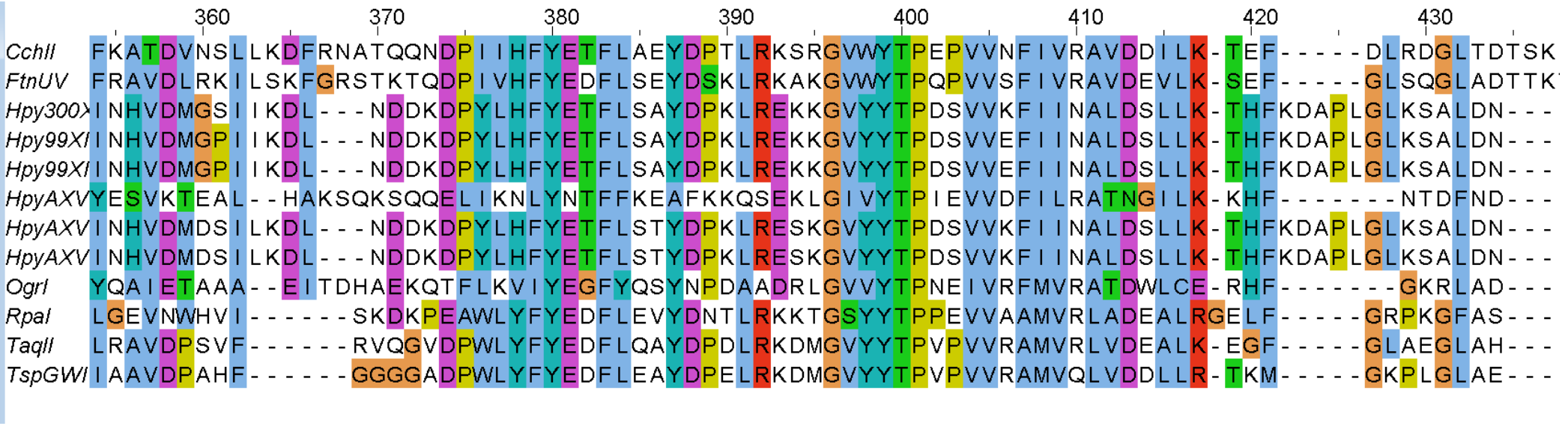
Файл mtases-12.fasta



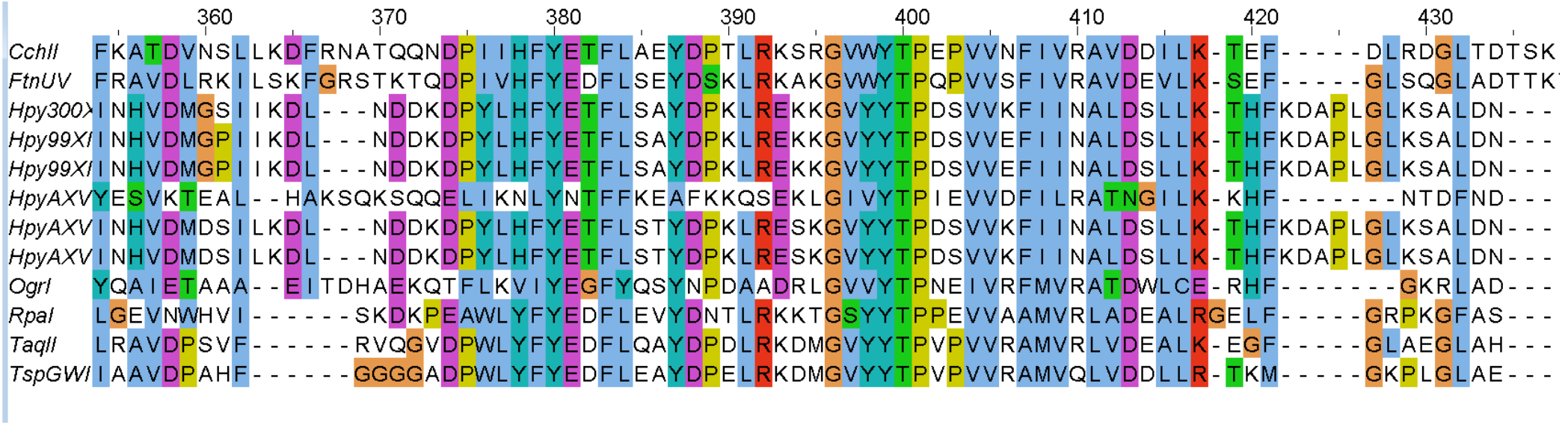
Участок 1


190
200
210
220
230
240
250
260

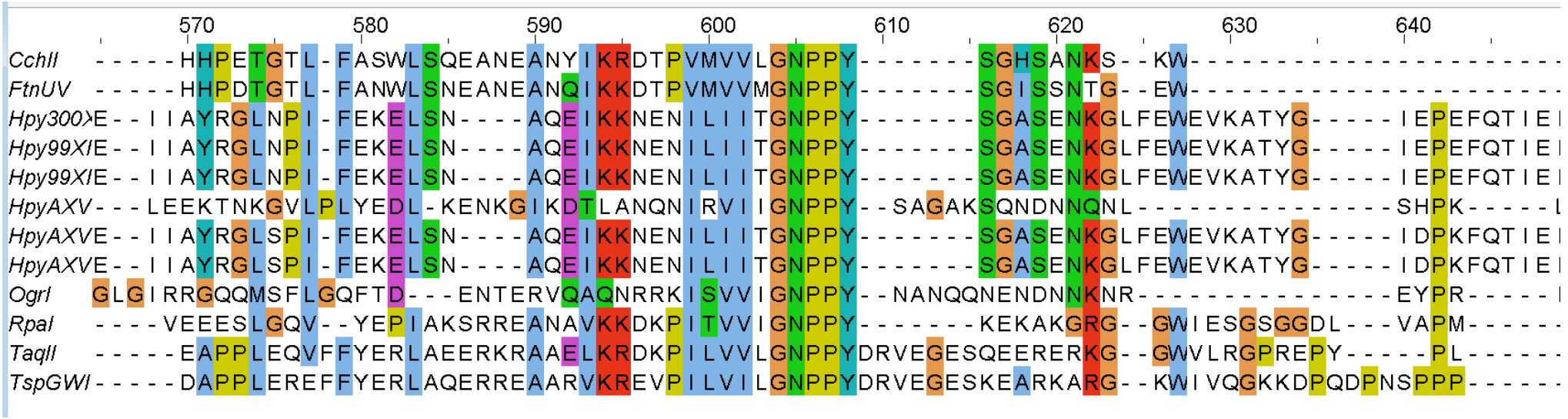

190
200
210
220
230
240
250
260


360
370
380
390
400
410
420
430
CchII FKATDVNSLLKDFRNATQQNDPIIHFYETFLAEYDPTLRKSRGVWYTPEPVVNFIVRAVDDILK-TEF-----DLRDGLTDTSK
FtnUV FRAVDLRKILSKFGRSTKTQDPIVHFYEDFLSEYDSKLRKAKGVWYTPQPVVSFIVRAVDEVLK-SEF-----GLSQGLADTTK
Hpy300X INHVDMGSIIKDL---NDDKDPYLHFYETFLSAYDPKLREKKGVYYTPDSVVKFIINALDSLLK-THFKDAPLGLKSALDN---
Hpy99XI INHVDMGPIIKDL---NDDKDPYLHFYETFLSAYDPKLREKKGVYYTPDSVVEFIINALDSLLK-THFKDAPLGLKSALDN---
Hpy99XI INHVDMGPIIKDL---NDDKDPYLHFYETFLSAYDPKLREKKGVYYTPDSVVEFIINALDSLLK-THFKDAPLGLKSALDN---
HpyAXV YESVKTEAL--HAKSQKSQQELIKNLYNTFFKEAFKKQSEKLGIVYTPIEVVDFILRATNGILK-KHF-------NTDFND---
HpyAXV INHVDMDSILKDL---NDDKDPYLHFYETFLSTYDPKLRESKGVYYTPDSVVKFIINALDSLLK-THFKDAPLGLKSALDN---
HpyAXV INHVDMDSILKDL---NDDKDPYLHFYETFLSTYDPKLRESKGVYYTPDSVVKFIINALDSLLK-THFKDAPLGLKSALDN---
OgrI YQAIETAAA--EITDHAEKQTFLKVIYEGFYQSYNPDAADRLGVVYTPNEIVRFMVRATDWLCE-RHF-------GKRLAD---
RpaI LGEVNWHVI------SKDKPEAWLYFYEDFLEVYDNTLRKKTGSYYTPPEVVAAMVRLADEALRGELF-----GRPKGFAS---
TaqII LRAVDPSVF------RVQGVDPWLYFYEDFLQAYDPDLRKDMGVYYTPVPVVRAMVRLVDEALK-EGF-----GLAEGLAH---
TspGWI IAAVDPAHF------GGGGADPWLYFYEDFLEAYDPELRKDMGVYYTPVPVVRAMVQLVDDLLR-TKM-----GKPLGLAE---

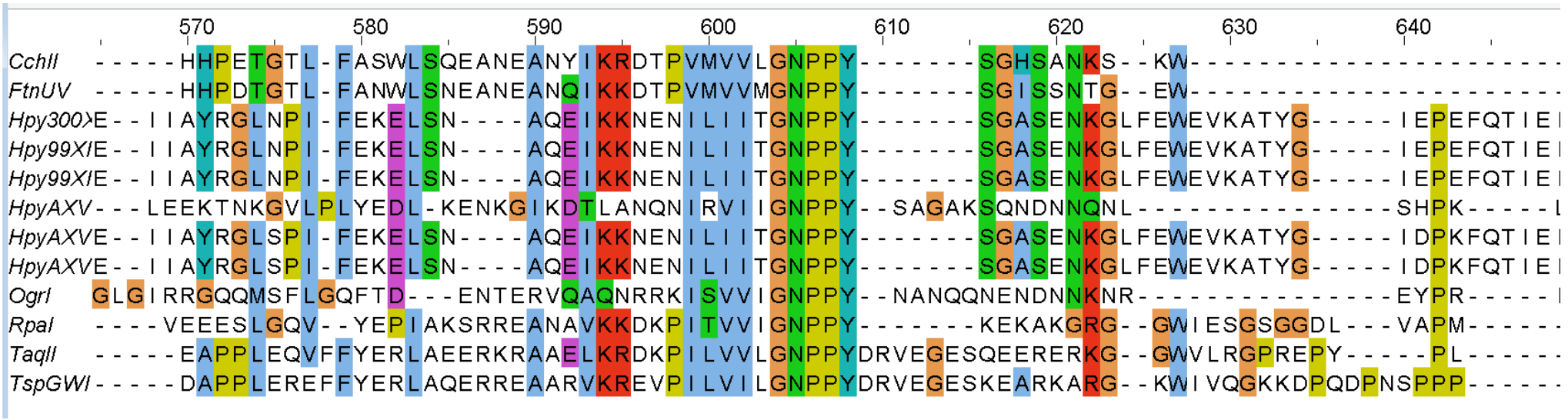
Участок 2

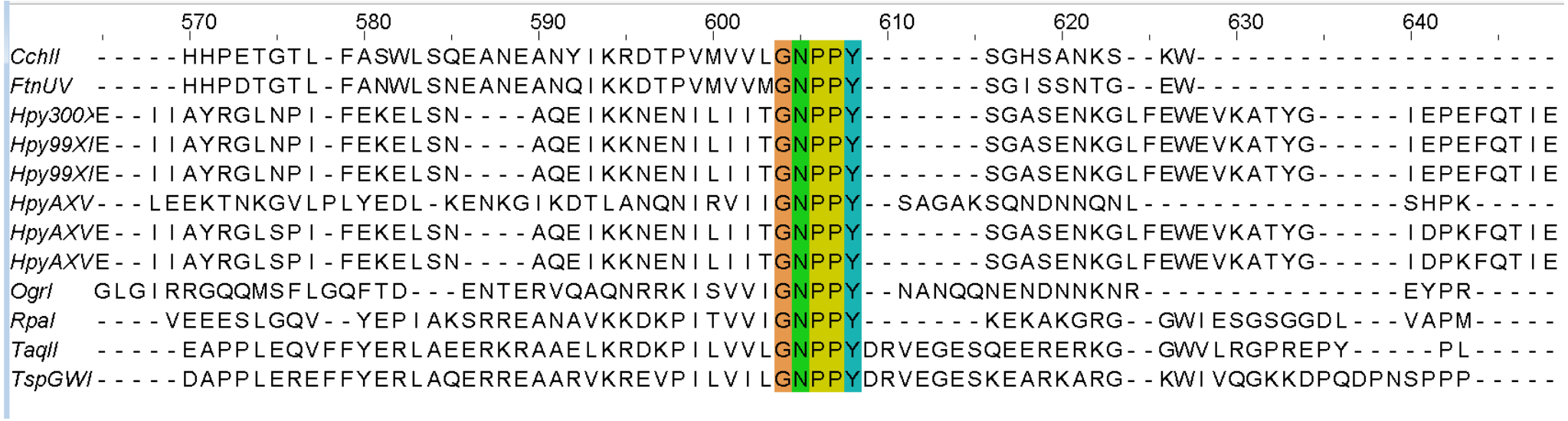


360 370 380 390 400 410 420 430
CchII FKATDVNSLLKDFRNATQQNDPIIHFYETFLAEYDPTLRKSRGVWYTPEPVVNFIVRAVDDILK-TEF-----DLRDGLTDTSK
FtnUV FRAVDLRKILSKFGRSTKTQDPIVHFYEDFLSEYDSKLRKAKGVWYTPQPVVSFIVRAVDEVLK-SEF-----GLSQGLADTTK
Hpy300X INHVDMGSIIKDL---NDDKDPYLHFYETFLSAYDPKLREKKGVYYTPDSVVKFIINALDSLLK-THFKDAPLGLKSALDN---
Hpy99XI INHVDMGPIIKDL---NDDKDPYLHFYETFLSAYDPKLREKKGVYYTPDSVVEFIINALDSLLK-THFKDAPLGLKSALDN---
Hpy99XI INHVDMGPIIKDL---NDDKDPYLHFYETFLSAYDPKLREKKGVYYTPDSVVEFIINALDSLLK-THFKDAPLGLKSALDN---
HpyAXV YESVKTEAL--HAKSQKSQQELIKNLYNTFFKEAFKKQSEKLGIVYTPIEVVDFILRATNGILK-KHF-------NTDFND---
HpyAXV INHVDMDSILKDL---NDDKDPYLHFYETFLSTYDPKLRESKGVYYTPDSVVKFIINALDSLLK-THFKDAPLGLKSALDN---
HpyAXV INHVDMDSILKDL---NDDKDPYLHFYETFLSTYDPKLRESKGVYYTPDSVVKFIINALDSLLK-THFKDAPLGLKSALDN---
OgrI YQAIETAAA--EITDHAEKQTFLKVIYEGFYQSYNPDAADRLGVVYTPNEIVRFMVRATDWLCE-RHF-------GKRLAD---
RpaI LGEVNWHVI------SKDKPEAWLYFYEDFLEVYDNTLRKKTGSYYTPPEVVAAMVRLADEALRGELF-----GRPKGFAS---
TaqII LRAVDPSVF------RVQGVDPWLYFYEDFLQAYDPDLRKDMGVYYTPVPVVRAMVRLVDEALK-EGF-----GLAEGLAH---
TspGWI IAAVDPAHF------GGGGADPWLYFYEDFLEAYDPELRKDMGVYYTPVPVVRAMVQLVDDLLR-TKM-----GKPLGLAE---


570
580
590
600
610
620
630
640
CchII -----HHPETGTL-FASWLSQEANEANYIKRDTPVMVVLGNPPY-------SGHSANKS--KW----------
FtnUV -----HHPDTGTL-FANWLSNEANEANQIKKDTPVMVVMGNPPY-------SGISSNTG--EW----------
Hpy300X E--IIAYRGLNPI-FEKELSN----AQEIKKNENILIITGNPPY-------SGASENKGLFEWEVKATYG-----IEPEFQTIEI
Hpy99XI E--IIAYRGLNPI-FEKELSN----AQEIKKNENILIITGNPPY-------SGASENKGLFEWEVKATYG-----IEPEFQTIEI
Hpy99XI E--IIAYRGLNPI-FEKELSN----AQEIKKNENILIITGNPPY-------SGASENKGLFEWEVKATYG-----IEPEFQTIEI
HpyAXV ---LEEKTNKGVLPLYEDL-KENKGIKDTLANQNIRVIIGNPPY--SAGAKSQNDNNQNL---------------SHPK-----L
HpyAXVE--IIAYRGLSPI-FEKELSN----AQEIKKNENILIITGNPPY-------SGASENKGLFEWEVKATYG-----IDPKFQTIEI
HpyAXVE--IIAYRGLSPI-FEKELSN----AQEIKKNENILIITGNPPY-------SGASENKGLFEWEVKATYG-----IDPKFQTIEI
Ogrl GLGIRRGQQMSFLGQFTD---ENTERVQAQNRRKISVVIGNPPY--NANQQNENDNNKNR---------------EYPR-----I
RpaI ----VEEESLGQV--YEPIAKSRREANAVKKDKPITVVIGNPPY-------KEKAKGRG--GWIESGSGGDL---VAPM------
TaqII -----EAPPLEQVFFYERLAEERKRAAELKRDKPILVVLGNPPYDRVEGESQEERERKG--GWVLRGPREPY-----PL------
TspGWI -----DAPPLEREFFYERLAQERREAARVKREVPILVILGNPPYDRVEGESKEARKARG--KWIVQGKKDPQDPNSPPP------

Участок 3


570 580 590 600 610 620 630 640
CchII -----HHPETGTL-FASWLSQEANEANYIKRDTPVMVVLGNPPY-------SGHSANKS--KW----------
FtnUV -----HHPDTGTL-FANWLSNEANEANQIKKDTPVMVVMGNPPY-------SGISSNTG--EW----------
Hpy300X E--IIAYRGLNPI-FEKELSN----AQEIKKNENILIITGNPPY-------SGASENKGLFEWEVKATYG-----IEPEFQTIEI
Hpy99XI E--IIAYRGLNPI-FEKELSN----AQEIKKNENILIITGNPPY-------SGASENKGLFEWEVKATYG-----IEPEFQTIEI
Hpy99XI E--IIAYRGLNPI-FEKELSN----AQEIKKNENILIITGNPPY-------SGASENKGLFEWEVKATYG-----IEPEFQTIEI
HpyAXV ---LEEKTNKGVLPLYEDL-KENKGIKDTLANQNIRVIIGNPPY--SAGAKSQNDNNQNL----------SHPK-----I
HpyAXV E--IIAYRGLSPI-FEKELSN----AQEIKKNENILIITGNPPY-------SGASENKGLFEWEVKATYG-----IDPKFQTIEI
HpyAXV E--IIAYRGLSPI-FEKELSN----AQEIKKNENILIITGNPPY-------SGASENKGLFEWEVKATYG-----IDPKFQTIEI
Ogrl GLGIRRGQQMSFLGQFTD---ENTERVQAQNRRKISVVIGNPPY--NANQQNENDNNKNR----------EYPR-----I
RpaI ----VEEESLGQV--YEPIAKSRREANAVKKDKPITVVIGNPPY-------KEKAKGRG--GWIESGSGGDL---VAPM------
TaqII -----EAPPLEQVFFYERLAEERKRAAELKRDKPILVVLGNPPYDRVEGESQEERERKG--GWVLRGPREPY-----PL------
TspGWI -----DAPPLEREFFYERLAQERREAARVKREVPILVILGNPPYDRVEGESKEARKARG--KWIVQGKKDPQDPNSPPP------


570 580 590 600 610 620 630 640
CchII -----HHPETGTL-FASWLSQEANEANYIKRDTPVMVVLGNPPY-------SGHSANKS--KW----------
FtnUV -----HHPDTGTL-FANWLSNEANEANQIKKDTPVMVVMGNPPY-------SGISSNTG--EW----------
Hpy300X E--IIAYRGLNPI-FEKELSN----AQEIKKNENILIITGNPPY-------SGASENKGLFEWEVKATYG-----IEPEFQTIE
Hpy99XI E--IIAYRGLNPI-FEKELSN----AQEIKKNENILIITGNPPY-------SGASENKGLFEWEVKATYG-----IEPEFQTIE
Hpy99XI E--IIAYRGLNPI-FEKELSN----AQEIKKNENILIITGNPPY-------SGASENKGLFEWEVKATYG-----IEPEFQTIE
HpyAXV ---LEEKTNKGVLPLYEDL-KENKGIKDTLANQNIRVIIGNPPY--SAGAKSQNDNNQNL----------SHPK-----
HpyAXV E--IIAYRGLSPI-FEKELSN----AQEIKKNENILIITGNPPY-------SGASENKGLFEWEVKATYG-----IDPKFQTIE
HpyAXV E--IIAYRGLSPI-FEKELSN----AQEIKKNENILIITGNPPY-------SGASENKGLFEWEVKATYG-----IDPKFQTIE
Ogrl GLGIRRGQQMSFLGQFTD---ENTERVQAQNRRKISVVIGNPPY--NANQQNENDNNKNR----------EYPR-----
RpaI ----VEEESLGQV--YEPIAKSRREANAVKKDKPITVVIGNPPY-------KEKAKGRG--GWIESGSGGDL---VAPM-----
TaqII -----EAPPLEQVFFYERLAEERKRAAELKRDKPILVVLGNPPYDRVEGESQEERERKG--GWVLRGPREPY-----PL-----
TspGWI -----DAPPLEREFFYERLAQERREAARVKREVPILVILGNPPYDRVEGESKEARKARG--KWIVQGKKDPQDPNSPPP-----

Составление паттерна и поиск по паттерну

fuzzpro, fuzznuc, Jalview

Поиск по паттерну в Jalview

- Составьте паттерны найденных мотивов с высоким IC. Участок 3:
  - В примере select => find => **GNPPY** => Find all
    12 находок. Лишних нет
- Участок 2 ослабим условие – разрешим замены на остатки с похожими физико-химическими свойствами!
  - **[GA].{2}[FY][TS]P…[VLIM]**

  [GA] = либо G, либо A; Точка . = любая буква; {2} = повторяется 2 раза
  12 находок. Лишних нет
- Участок 1:
  - **L.{30,34}L**
    {30,34} = повторяется от 30 до 34 раз
    509 находок – низкое IC, мотив не пригоден для поиска

Паттерны

Запись мотива в белке в виде регулярного выражения для ряда программ

Правила записи:
https://myhits.sib.swiss/cgi-bin/help?doc=pattern.html

Пример паттерна

< A-x-[ST](2)-x(0,1)-V

Отличия от паттерна Jalview:

1) Разделители чёрточка –
2) x вместо точки обозначает любую букву
3) число повторений в скобках () а не {}

Поиск паттернов в наборах белков реализован

- В базе данных MyHits https://myhits.sib.swiss/cgi-bin/pattern_search в SwissProt и наборах протеомов.
- Содержит также коллекцию паттернов профилей и др.
- Есть поиск известных паттернов во входном белке.
- Хорошо – понятно - организованная база данных в Лозанне.
- Отстаёт по объёму коллекций от лидеров, но обгоняет их по аккуратности
- Для учебных целей (и не только) подходит

Банк ProSite. https://prosite.expasy.org/

Коллекция белковых семейств и доменов.

Аннотации эволюционных доменов.

Мотивы: функциональные участки и «подписи» семейств белков в виде паттернов и HMM-профилей. pfTools

Интерфейс (средства поиска, средства сохранения выравниваний и т. д.)

Можно:

1. Искать мотивы из коллекции ProSite в своем белке.
2. Искать свой мотив в коллекции последовательностей ProSite.
3. Искать свой мотив в своем белке или белках.

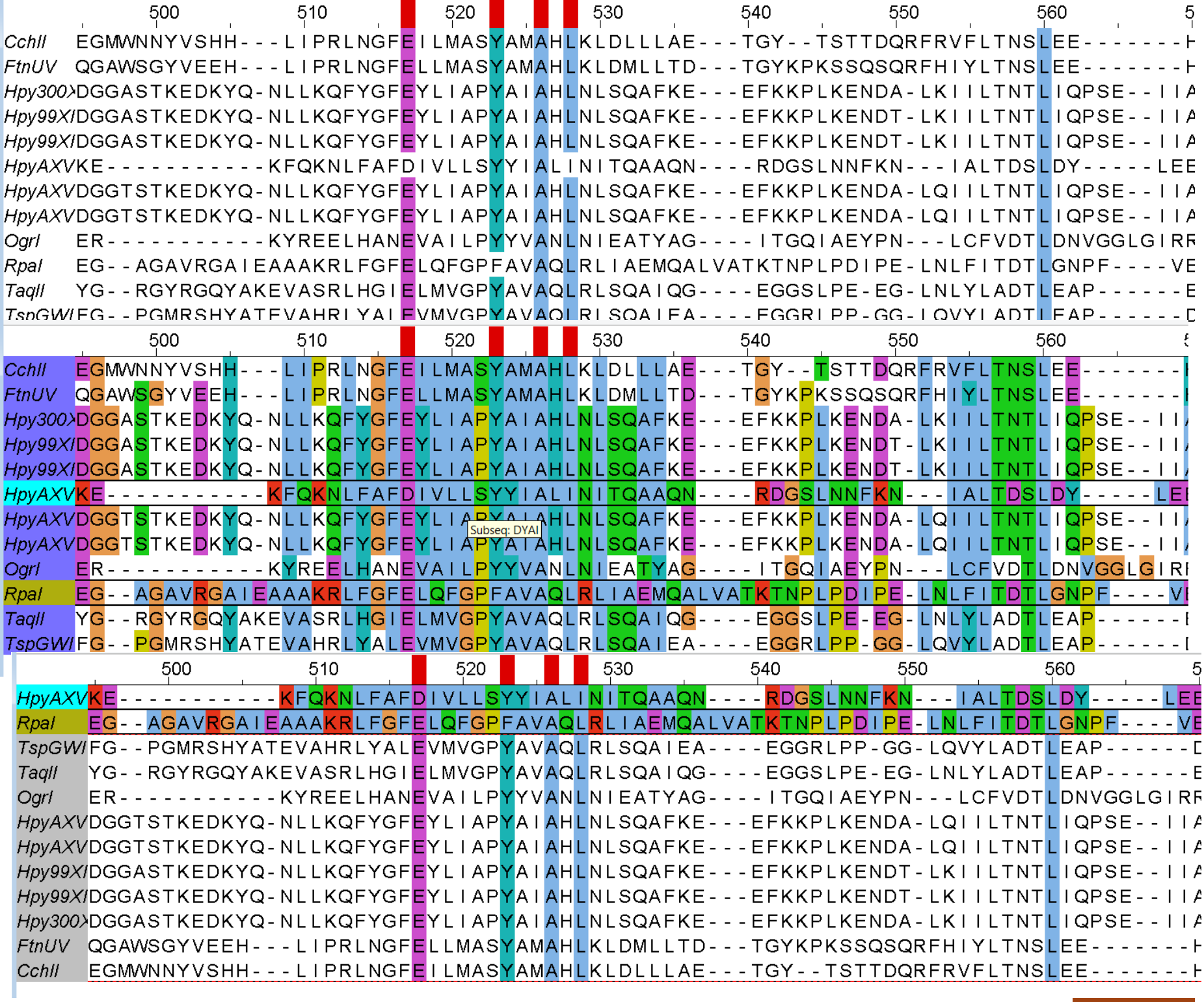
Поиск паттернов в наборах белков реализован

- В пакете EMBOSS.
- программа fuzznuc для поиска паттернов в нуклеотидных последовательностях
- fuzzpro для поиска паттернов в белковых последовательностях.

На вход — паттерн и последовательность, на выходе — позиция и вес найденных совпадений

Как найти консервативные мотивы в Jalview

- Далеко не всегда найдутся 100% консервативные мотивы.
- В этом случае можно понизить порог identity threshold
- Увидеть мотив с высоким IC.
- Отсортировать по позициям мотива:
  - Выделить колонки
  - Select => make groups for selection
  - Calculate => Sort => By groups
- Посмотреть на последовательности, в которых не найден мотив.
  - Может мотив можно ослабить
  - Может ошибка в выравнивании
  - Может этого мотива нет в последовательностях и наличие мотива – объективный признак, разделяющий последовательности на две группы.


500
510
520
530
540
550
560
CchII EGMWNNYVSHH---LIPRLNGFEILMASYAMAHLKLDLLLAE---TGY--TSTTDQRFRVFLTNSLEE-------H
FtnUV QGAWSGYVEEH---LIPRLNGFELLMASYAMAHLKLDMLLTD---TGYKPKSSQSQRFHIYLTNSLEE-------H
Hpy300X DGGASTKEDKYQ-NLLKQFYGFEYLIAPYAIAHLNLSQAFKE---EFKKPLKENDA-LKIILTNTLIQPSE--IIA
Hpy99XI DGGASTKEDKYQ-NLLKQFYGFEYLIAPYAIAHLNLSQAFKE---EFKKPLKENDT-LKIILTNTLIQPSE--IIA
Hpy99XI DGGASTKEDKYQ-NLLKQFYGFEYLIAPYAIAHLNLSQAFKE---EFKKPLKENDT-LKIILTNTLIQPSE--IIA
HpyAXV KE----------KFQKNLFAFDIVLLSYYIALINITQAAQN----RDGSLNNFKN---IALTDSLDY-----LEE
HpyAXV DGGTSTKEDKYQ-NLLKQFYGFEYLIAPYAIAHLNLSQAFKE---EFKKPLKENDA-LQIILTNTLIQPSE--IIA
HpyAXV DGGTSTKEDKYQ-NLLKQFYGFEYLIAPYAIAHLNLSQAFKE---EFKKPLKENDA-LQIILTNTLIQPSE--IIA
Ogrl ER----------KYREELHANEVAILPYYVANLNIEATYAG----ITGQIAEYPN---LCFVDTLDNVGGLGIRR
Rpal EG--AGAVRGAIEAAAKRLFGFELQFGPFAVAQLRLIAEMQALVATKTNPLPDIPE-LNLFITDTLGNPF----VE
Taqll YG--RGYRGQYAKEVASRLHGIELMVGPYAVAQLRLSQAIQG----EGGSLPE-EG-LNLYLADTLEAP------E
TspGWI FG--PGMRSHYATEVAHRLYALEVMVGPYAVAQLRLSQAIEA----EGGRLPP-GG-LQVYLADTLEAP------D
Subseq: DYAI

Задание 2. Найти мотивы специфичные для клады филогенетического дерева в Jalview

- Построить филогенетическое дерево
  - Calculate => Calculate phylogenetic tree or PCA => NJ or Average distance = UPGMA
  - На рисунке дерева щелчком по верхней прямой разрезать ветку, отделяющую кладу. При этом разрежутся и другие ветки, т.к. вертикальная линия разреза разрезает ветви по всей высоте!
  - Соответственно разрезам в выравнивании выделятся группы, среди них нужная клада.
- Сортировка по группам: Calculate => Sort => By groups
- Выделите последовательности выбранной клады и сделайте из них выравнивание в отдельном окне:
  - Правок кнопкой на выделенных последовательностях
  - Selection => Output to text box => fasta
  - В новом окне => New windows
  - Предыдущее окно можно закрыть – оно больше не понадобится.
  - Edit => remove empty columns. Выравнивание клады готово

Задание 2. Найти мотивы специфичные для клады филогенетического дерева в Jalview

- Сохраните все окна в файле с Project
  - В самом внешнем окне File => Save Project
- В полученном выравнивании клады найдите мотивы.
- Поверьте каждый на то, что он специфичен для клады. Т.е. во всем выравнивании находятся по нему последовательности клады и ничего больше!

Если получилось, то вы нашли объективный признак, подтверждающий правильность соответствующей ветви филогенетического дерева.

IIc. PSSM и PSI BLAST для белков

С.А.Спирин, А.С.Ершова

Вспомним PWM, вес и информационное содержание

TTATGCC
ATCTTCA
GTATTAA

выравнивание

	1	2	3	4	5	6	7
G	0.26	-1.3	-1.3	-1.3	0.26	-1.3	-1.3
A	0.26	-1.3	0.74	-1.3	-1.3	0.26	0.74
T	0.26	1.18	-1.3	1.18	0.74	-1.3	-1.3
C	-1.3	-1.3	0.26	-1.3	-1.3	0.74	0.26

PWM для данного выравнивания

$$I_k = \sum_i f_{ki} \log_2 \frac{f_{ki}}{p_i}$$

$$I = \sum_k I_k$$

Информационное содержание (I) позволяет понять, как много похожих на мотив последовательностей мы найдем в наших данных по случайным причинам.

Элементы PWM: S_{ki} для основания i в позиции k,

p_i — фоновая частота основания i

f_{ki} — частота основания i в позиции k

(с учётом псевдоотсчётов)

λ — любое число (для удобства)

$$S_{ki} = \frac{1}{\lambda} \log \frac{f_{ki}}{p_i}$$

Применение PWM

Приложив позиционную весовую матрицу (PWM) к последовательности той же длины, можно понять, содержит ли последовательность сигнал, описываемый этой PWM.

Чем выше вес, тем более вероятно, что последовательность содержит сигнал.

можно искать вероятные вхождения мотива в длинную последовательность (например, геном), считая вес всех возможных отрезков нужной длины: где вес выше порога, там предсказывается мотив. Выбор порога — отдельная задача.

PSSM — position-specific scoring matrix

По смыслу PSSM — это то же, что PWM, но термин PWM используется для мотивов в ДНК, а PSSM и для мотивов в белках и для описания семейств родственных белков, т.е. описания длинных участков выравнивания

Можно использовать гэпы, учитываются как 21 буква. PSSM применяется так же, как PWM: если вес последовательности белка относительно PSSM выше порога, предсказывается принадлежность белка семейству.

Базовая идея — та же, что для PWM:

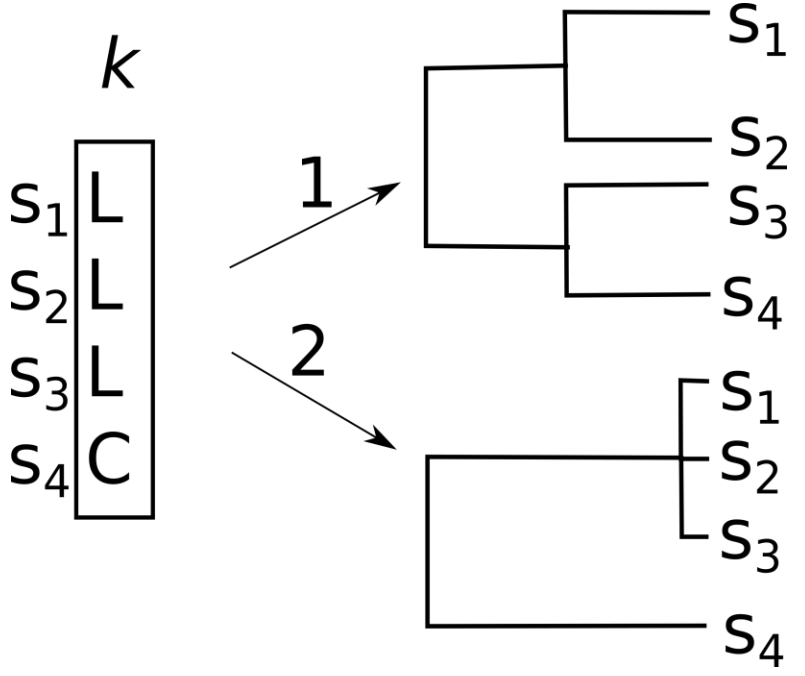
$$S_{ki} = \frac{1}{\lambda} \log \frac{f_{ki}}{p_i}$$

где S_{ki} — элемент позиционной весовой матрицы (вес буквы i в позиции k),

p_i — фоновая частота остатка i

f_{ki} — частота остатка i в позиции k (с учётом псевдоотсчётов)

В PSSM применяется взвешивание последовательностей (в отличие от PWM)



Колонка k выглядит так, что в ней предпочтительна буква L

Это может случиться из-за того, что много почти совпадающих последовательностей – случай 2

Поэтому вес от близкородственных последовательностей следует уменьшить так, чтобы он был близким к весу одной уникальной последовательности.

Это делается путем приписывания веса ω_s каждой последовательности s так, чтобы у последовательностей, имеющих много родственников, он был маленьким, а у «одиноких» последовательностей — большой.

Формула для частоты f_{ki} веса буквы i в колонке k выглядит так:

$$f_{ki} = \frac{\sum\limits_{s:a_{sk}=i} w_s + \psi_i}{\sum\limits_s w_s + \sum\limits_i \psi_i}$$

Каждая буква i в колонке k считается не за единицу, а за её вес ω_s
Здесь a_{sk} — буква последовательности s в позиции k , ψ_i — псевдоотсчёт для буквы i.

Внимание: слово «вес» имеет два разных значения

- Вес = Score, вес выравнивания двух последовательностей или последовательности относительно профиля (PWM или PSSM или HMM), обычно обозначается S.
- Вес = Weight, вес последовательности, используемый при построении PSSM по множественному выравниванию, обычно обозначается w.

Для белков имеются разные способы
(i) приписывания веса ω_s последовательности s
(ii) определения псевдоотсчётов ψ_i для буквы i.

Окончательная формула для элемента PSSM

$$S_{ki} = \frac{1}{\lambda} \log \frac{Q_{ki}}{p_i}$$

где S_{ki} — элемент PSSM (вес остатка i в позиции k),
Q_{ki} — ожидаемая частота остатка i в позиции k, с учетом весов последовательностей и псевдоотсчетов,
p_i — фоновая частота остатка i ,
λ — константа (для удобства)

Использование PSSM

PSSM можно «выравнять» с белковой последовательностью и получить вес, аналогично весу выравнивания двух последовательностей.

PSSM используется при поиске в банке данных программой PSI-BLAST и программами пакета MEME.

PSI-BLAST (Position-Specific Iterative BLAST) — разновидность BLASTP, использующий PSSM, благодаря чему он способен находить дальних родственников заданного белка.

Алгоритм PSI-BLAST

На входе — последовательность и порог по e-value, на выходе — набор найденных последовательностей и построенный по ним PSSM.

1. На первом этапе запускается обычный BLASTP входной последовательности против выбранного банка последовательностей
2. Для находок со значениями e-value лучше заданного порога строится множественное выравнивание.
3. Это выравнивание используется для получения PSSM.
4. На следующем шаге опять происходит запуск BLAST для исходной последовательности против того же банка последовательностей, но вместо матрицы замен остатков используется PSSM, полученная на предыдущем шаге.
5. Повторяем шаги 2-4, пока не перестанут добавляться новые последовательности.

Дополнительные возможности PSI-BLAST

Можно вручную включать/исключать последовательности, которые используются для построения PSSM

Можно использовать PSSM, созданную на основе поиска в одном банке, для поиска в другом банке.

III. Алгоритмы поиска мотивов в последовательностях

* MEME: Multiple Expectation Maximization for Motif Elicitation

Алгоритм MEME

1. Последовательно берем фрагмент заданной длины в каждой последовательности, ищем похожие фрагменты в других последовательностях, строим выравнивание. Берем базовые частоты букв из дополнения.
2. Для каждого выравнивания получаем PWM с максимальным весом, используя алгоритм EM (Expectation maximization)
3. Выбираем заданное число PWM с лучшим весом
4. Если задан поиск мотивов разной длины, то все заказанные длины перебираются
5. Программы пакета MEME строят и используют не PWM, а PFM - частотную матрицу. PWM строится переводом частот в логарифмы (описано в презентации к Л8)

Алгоритм EM (Expectation maximization)

- На входе выравнивание и PWM
- По очереди удаляем фрагмент из выравнивания, и заменяем его на лучший по PWM фрагмент в соответствующей последовательности
- Повторяем пока процесс не сходится
- Находим максимальный вес, записываем PWM с максимальным весом

E-value мотива, найденного с помощью MEME

- MEME улучшалась несколько раз
- В классическом варианте
  - Нужно одно число на выравнивание (аналог веса для BLAST). Это число – информационное содержание
  - E-value должно показывать мат.ожидание числа мотивов с тем же или большим IC, получаемых поиском MEME в случайном банке того же размера и состава
  - Хорошей математической теории, позволяющей быстро вычислить E-value нет.
  - Используют эвристические алгоритмы

Задача поиска мотивов *de novo* у белков

Интересна

1) для поиска мотивов в белках с общей функцией, но недостаточно схожих по последовательности для построения обоснованного выравнивания по всей длине

2) Для гомологичных последовательностей тоже полезна, так как может не только найти мотивы, но и для верификации выравнивания: стоят ли находки мотива в тех же колонках выравнивания

Работает так же, как для последовательностей нуклеиновых кислот

Ограничения MEME

1. Предположение о независимости позиций выравнивания
2. Находит только мотивы без гэпов
3. Предназначен для коротких последовательностей
4. Число входных последовательностей ограничено (<50)

3. Find Individual Motif Occurrences (FIMO)

1. FIMO ищет встречи каждого из входных мотивов по очереди, независимо друг от друга
2. Использует алгоритм динамического программирования
3. Вычисляет p-value для каждой находки.
4. Из-за проблемы множественного тестирования, p-value неправильно считать единственным показателем хорошей находки
5. FIMO сообщает и *p*-value, и, дополнительно, величину *q*-value, которая определяет порог false discovery rate (FDR) при котором *p*-value значимо

OXFORD
UNIVERSITY PRESS

FIMO output

From: **MEME Suite: tools for motif discovery and searching**

Nucleic Acids Res. 2009;37(suppl_2):W202-W208. doi:10.1093/nar/gkp335

Sequence Analysis with fimo

Pattern Name	Sequence Name	Start	Stop	Score	p-value	q-value	Matched Sequence
1	NP_418484.4|yjcB	281	298	21.2367	5.3e-09	0.00758	AATTGTGATATAGTTCAC
1	NP_418485.1|yjcC	149	132	21.2367	5.3e-09	0.00758	AATTGTGATATAGTTCAC
1	NP_418031.1|yiaJ	175	158	19.8034	3.86e-08	0.0173	AAGTGTGCCGTAGTTCAC
1	NP_418032.1|yiaK	26	43	19.8034	3.86e-08	0.0173	AAGTGTGCCGTAGTTCAC
1	NP_418535.1|proP	37	54	19.7078	4.3e-08	0.0173	ATGTGTGAAGTTGATCAC
1	NP_414666.1|gcd	126	143	19.6123	4.85e-08	0.0173	AATTGTGATGACGATCAC
1	NP_414667.4|hpt	80	63	19.6123	4.85e-08	0.0173	AATTGTGATGACGATCAC

Figure Legend:

IV Недопредставленность и перепредставленность слов (коротких последовательностей) и паттернов в геноме

Если слово значимо недопредставлено или перепредставлено, надо искать биологическую причину отбора, направленного на уменьшение числа слов, или на их увеличение.

Compositional Bias (CB) паттерна = (наблюдаемое число)/ожидаемое

(называют также контрастом)

CB >> 1 - слово перепредставлено

CB << 1 - слово недопредставлено

Вопрос как правильно вычислять «ожидаемое»

Метод «по бернулли» не всегда работает

- Такой способ подразумевает, что геном устроен как бернуллиевская случайная последовательность: появление следующей буквы не зависит от предыдущей.
- Это предположение нарушается.
- Пример. Число слов TA в большинстве геномов меньше, чем ожидаемое по бернулли. Поэтому число слов CTAG в геномах будет менее ожидаемого по бернулли!
- Как учесть этот эффект – учитывать частоты подслов!

K. Метод Карлина с соавт.

Учёт частот всех подслов, включая разрывные

$$CB = F_o/F_{exp}$$

$$F_{exp} = \frac{\prod Fo(oddN)}{\prod Fo(evenN)}$$

CB(GATC) =

- GATN, GANC, GNTC, NATC — 1N
- GANN, GNNC, NNTC, NATN, NANC, GNTN — 2N
- GNNN, NANN, NNTN, NNNC — 3N

См. Rusinov et al., Comparison of Methods of Detection of Exceptional Sequences in Prokaryotic Genomes, Biochemistry (Mosc.), 2018

ММ. Марковская цепь максимального порядка

$СВ = F_o/F_{exp}$

F_o – observed (наблюдаемое число слов)

F_{exp} – expected (ожидаемое число слов

$$Fexp\ (w1\ldots wm) = Fo\ (w1\ldots wm{-}1)\ Fo\ (w2\ldots wm)/\ Fo\ (w2\ldots wm{-}1).$$

Можно объяснить так (для CTAG). Перед словом TA с частотой $F_0(CTA)/F_0(TA)$ стоит буква C, с частотой $F_0(TAG)/F_0(TA)$ после TA стоит буква G. В предположении независимости имеем:

$Fo(CTAG) = F_0(CTA)/F_0(TA)* F_0(TA) * F_0(TAG))/F_0(TA) =$

$= (F_0(CTA) * F_0(TAG))/F_0(TA)$

В этой модели можно вычислить pValue отличия наблюдаемого от ожидаемого. См Gelfand, Koonin, NAR, 1997

КОНЕЦ ПРЕЗЕНТАЦИИ

Gibbs Sampling

- Первый шаг такой же, как в MEME: выбор выравнивания A из случайных фрагментов
- Шаг состоит в удалении одного фрагмента и замене его случайным фрагментом из той же последовательности => новое выравнивание B
- Если I(B) > I(A), то берем B
- Если I(B) < I(A), то с вероятностью $P = \exp[\,(I(B) - I(A))\,/\,T\,]$

  берем B, иначе оставляем A
- В начале “температура” T большая => почти все замены на худшее выравнивание B принимаются; с каждым шагом температура понижается, так что все более жесткие условия на то, чтобы взять B.
- “Тепловой отжиг” (Как в ПЦР☺)

Есть и другие алгоритмы

- У А.А.Миронова с соавт
- ChiPMunk (В.Макеев, И.Кулаковский с соват.) (https://opera.autosome.ru/chipmunk/discovery)

- И др (см. https://molbiol-tools.ca/DNA_Motifs.htm)

RnaP субъединицы alpha и sigma-70

Хотел проделать то же, но не успел справиться. Консервативные мотивы нашел, но не знаю их связи с функциями белков

ated from 656 sequences (Sierro et al. 2008).

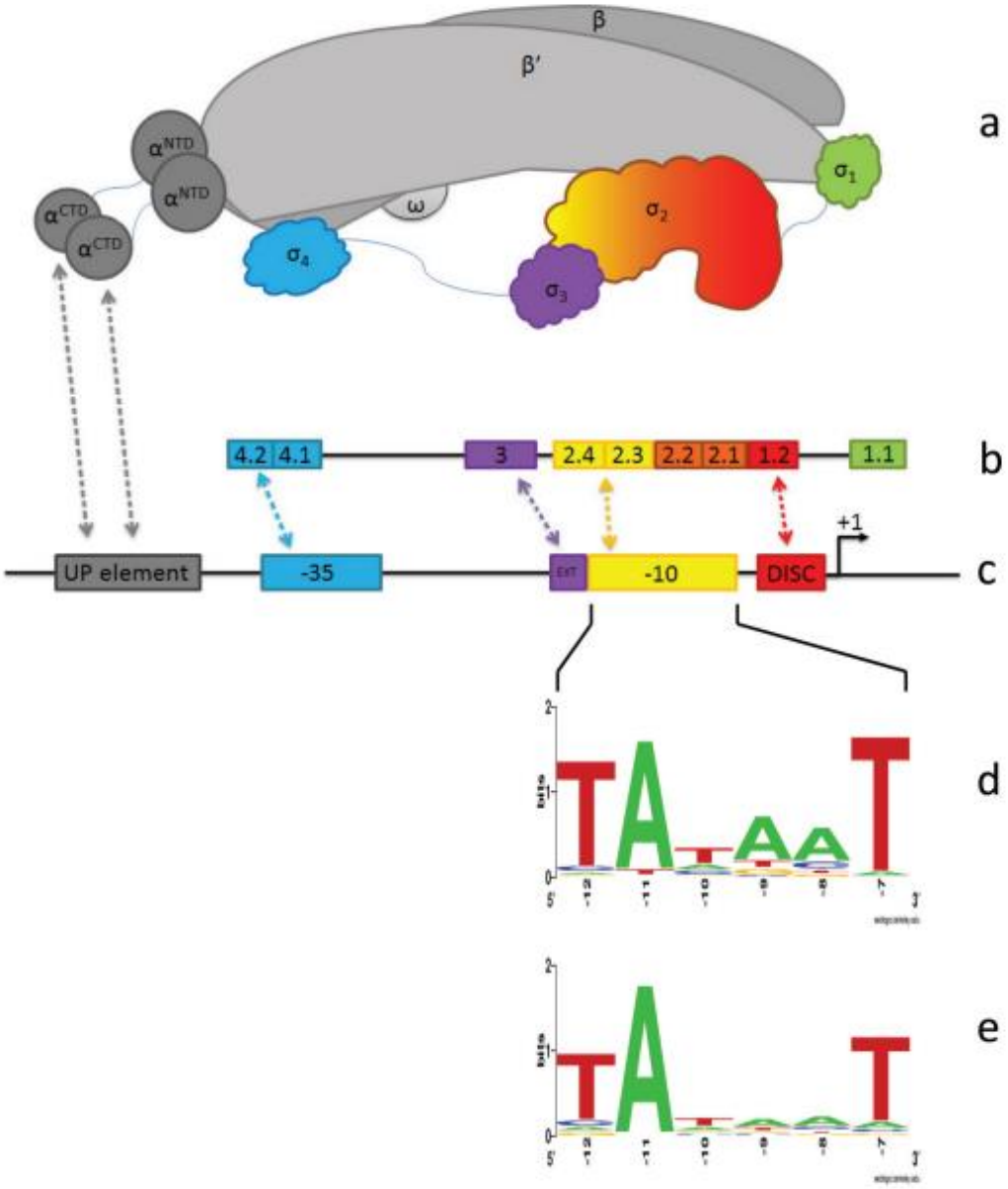


Fig. 1. Schematic representation of bacterial RNAP (RNA polymerase) holoenzyme, factor structure, and representative promoter structure. (a) Subunit architecture of holoenzyme (core subunits presented in various shades of grey); represented as 4 domains colour-coded according to (b) regions and subregions of contiguous amino acid residues in , and indicating typical interactions with (c) a representative 70 promoter with key elements indicated (EXT, extended –10 element; DISC, discriminator sequence). (d) Sequence logo (Schneider and Stephens 1990) for the Escherichia coli primary factor –10 element generated from 950 sequences (Gama-Castro et al. 2015). (e) Sequence logo for Bacillus subtilis primary factor –10 element generated from 656 sequences (Sierro et al. 2008).

Davis et al., The essential activities of the bacterial sigma factor, 2017

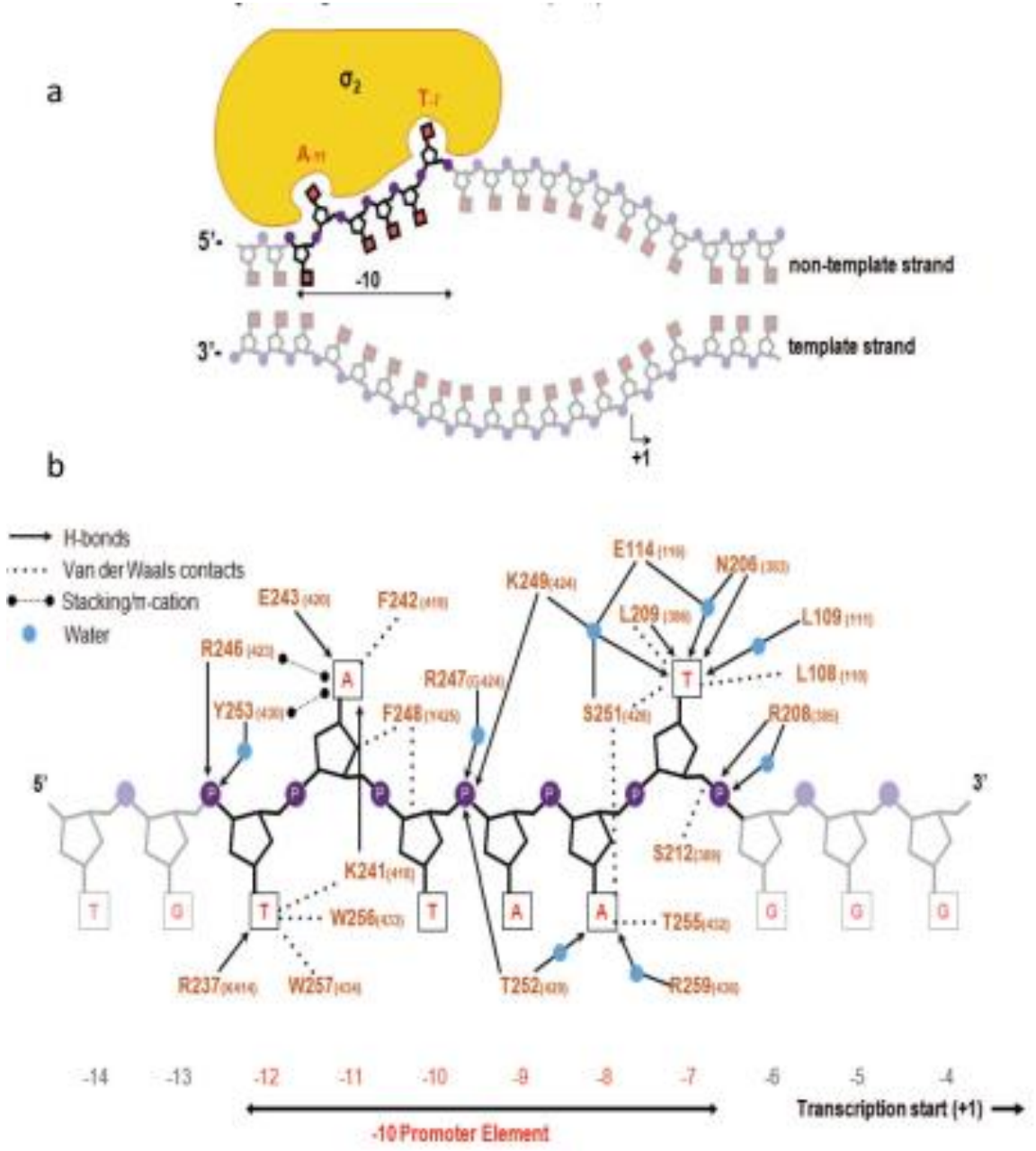


Fig. 3. Interaction patterns between Thermus aquaticus A 2 amino acid residues and –10 promoter element nucleotides. (a) Schematic diagram of open complex (represented in duplex DNA context) with 2 flipped out bases captured by domain 2. (b) Redrawing of inferred interactions between key amino acids (Escherichia coli 70 numbering in brackets) and the –10 element non-template strand as deduced and reported by Feklistov and Darst (2011).

Davis et al., The essential activities of the bacterial sigma factor, 2017

Содержание

1. Что было про сигналы в ДНК. Повторение
2. Мотивы в белках
   a. Jalview
   b. Fuzzpro
   c. PSSM и psiBLAST
3. Как находить новые мотивы в белках. MEME
4. Недопредставленность слов