

Сигналы и мотивы -3

Мотивы в последовательностях белков

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

I. Сигналы в ДНК. Что необходимо и достаточно знать для зачёта темы

- 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 поиска и описания 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 белках

Что такое?

короткие консервативные
последовательности в гомологичных белках

Консервативные – значит важные:

активные центры белков;

участки, связывающие лиганды;

Участки белок-белкового взаимодействия;

И др.

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

ДНК метилтрансферазы (Мтазы)

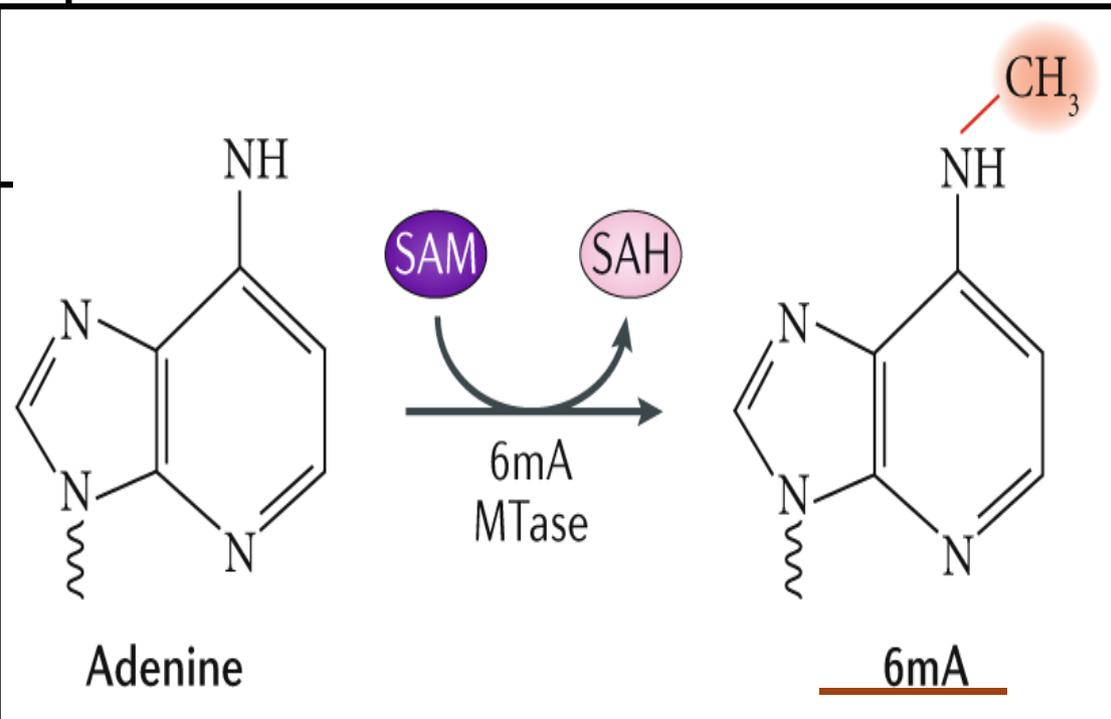
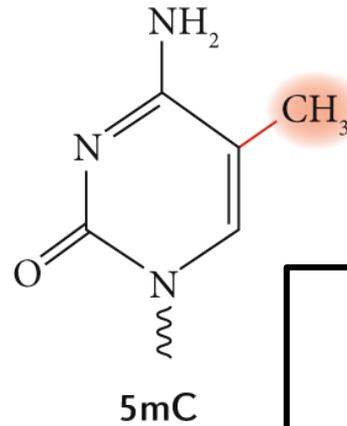
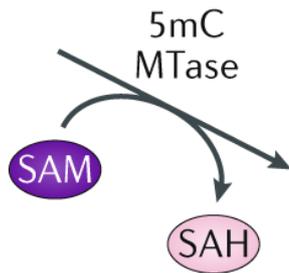
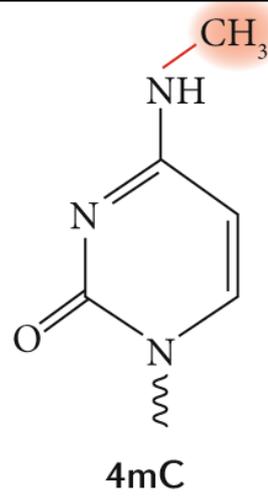
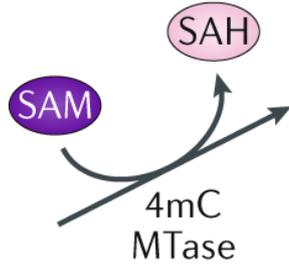
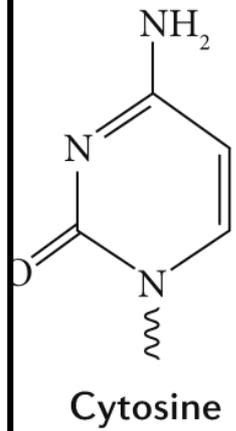
1. Узнают короткую последовательность ДНК (например, GATC)
 1. Узнаваемая последовательность может различаться даже у гомологичных Мтаз
2. Метилируют определенное основание ДНК в сайте или рядом.
3. Три вида метилирования:
 1. 5mC метилируется углерод цитозина в положении 5
 2. 4mC метилируется азот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

Ожидаем консервативность

3. Активного центра, но не у всех Мтаз
4. Участка связывания SAM

SAM S-adenosyl-L-methionine

SAH S-adenosyl-homocysteine



Найдём мотивы у МТаз

И опишем их паттернами используя
Jalview

Две задачи, при наличии выравнивания

- Задача 1: найти функционально важные участки белков
- Важные – скорее всего консервативные!
- Задача 2: для клады – множества последовательностей одной ветви филогенетического дерева - найдите мотивы у клады, которые отсутствуют в других последовательностях.

Это обязательное задание по Jalview. Начнем делать его на занятии. Начну с демонстрации выполнения на простом примере

Переход в окно Jalview

Выравнивание 12 Мтаз.

Файл `mtases-12.fasta`

Задача 1. Как найти консервативные мотивы в Jalview

- Выровнять последовательности. Посмотрите на выравнивание чтобы оценить его качество.
- Запишите сколько последовательностей
- Удалить идентичные или очень похожие последовательности
 - Edit => remove redundancy => 100% или меньше, например, 90%
- Запишите сколько последовательностей осталось
- Color
 - Above identity threshold 100%!
 - Изолированные консервативные колонки пока не рассматриваем – могут быть результатом подгонки алгоритма выравнивания
 - Вплотную или рядом расположенные консервативные позиции считаем мотивом
- Оцените “на глаз” информационное содержание (IC) – насколько далек от случайного совпадения найденный мотив. Упорядочите их по IC

190 200 210 220 230 240 250 260
 AFADLMQDFGGYEGQTIGS---ASKLSKMMAAKARLLADVLEKALD-GYSDENNGAIDEASNTLYDQLKGF RDV---
 QFENLIKDFCTYIGQTI RS---PKKLAEMMAGKARLLQNTLERALEQDIADDDNTE LNE-----QYETFKDI---
 DLIELFRGFFNHEAAPITN---AKDFANALSAPTRYLKDAL-----IQYQKDMQVSS-----IFNNFKEY---
 DLIELFRGFFNHEAAPITN---AKDFANALSAPTRYLKDAL-----IAYQKDDQVSS-----IFKNFKEY---
 DLIELFRGFFNHEAAPITN---AKDFANALSAPTRYLKDAL-----IAYQKDDQVSS-----IFKNFKEY---
 ---LADAMYNVMPKLGDRNYWENFTKKTGN IARTLNNRL-----KIFDKNPEFFH---G---FLDSLREN---
 DLIELFKSFFNHEAAPITN---AKDFATHLSPRTKY LKDAL-----IKYQEKAQVSS-----IFNNFKEY---
 DLIELFKSFFNHEAAPITN---AKDFATHLSPRTKY LKDAL-----IKYQEKAQVSS-----IFNNFKEY---
 DLNRLLVAFFDWQPAVIGEWRS AVQQFRVELPAI LGHLRERI-----DKAYDDNEAFTAKATA---FLQHARET---
 TLAPLIATFLQWSP IAPKS---AKALAQVSARLCRLLRDEV---IE-QLELGSAG-LTE-----LAKDWRHL---
 ALRELFDFLNWRPLVPRN---PQELARFLAPLARFLREAV---LE-EVRENPNGELAR-----LREEWRKN---
 RVEELL LDFLHWQPLVPKN---PQELARFLAPL TRFLREAV---VE-ALREDPEGR LAH-----LYREWAGDPASC

Участок 1

190 200 210 220 230 240 250 260
 AFADLMQDFGGYEGQTIGS - - - - ASKLSKMMAAKARLLADVLEKALD - GYSDENNGAIDEASNTLYDQLKGFRDVG - - - -
 QFENLIKDFCTYIGQTI RS - - - - PKKLAEMMAGKARLLQNTLERALEQDIADDDNTE LNE - - - - - QYETFKDI - - - -
 DLIELFRGFFNHEAAPITN - - - - AKDFANALSAPTRYLKDAL - - - - - IQYQKDMQVSS - - - - - IFNNFKEY - - - -
 DLIELFRGFFNHEAAPITN - - - - AKDFANALSAPTRYLKDAL - - - - - IAYQKDDQVSS - - - - - IFKNFKEY - - - -
 DLIELFRGFFNHEAAPITN - - - - AKDFANALSAPTRYLKDAL - - - - - IAYQKDDQVSS - - - - - IFKNFKEY - - - -
 - - - - LADAMYNVMP TKLGD - RNYWENFTKKTGNIARTLNNRL - - - - - KIIFDKNPEFFH - - - - - G - - - - FLDSLREN - - - -
 DLIELFKSFFNHEAAPITN - - - - AKDFATHLSPR TKY LKDAL - - - - - IKYQEKAQVSS - - - - - IFNNFKEY - - - -
 DLIELFKSFFNHEAAPITN - - - - AKDFATHLSPR TKY LKDAL - - - - - IKYQEKAQVSS - - - - - IFNNFKEY - - - -
 DLNRLLVAFFDWQPAVIGEWRS AVQQFRVELPAI LGHLRERI - - - - - DKAYDDNEAFTAKATA - - - - - FLQHARET - - - -
 TLAPLIATFLQWSP IAPKS - - - - AKALAQVSARLCRLLRDEV - - - - - IE - QLELGSAG - LTE - - - - - LAKDWRHL - - - -
 ALRELFLDFLNWRPLVPRN - - - - PQELARFLAPLARFLREAV - - - - - LE - EVRENPN GELAR - - - - - LREEWRKN - - - -
 RVEELL LDFLHWQPLVPKN - - - - PQELARFLAPL TRFLREAV - - - - - VE - ALREDPEGRLAH - - - - - LYREWAGDPAS

190 200 210 220 230 240 250 260
 AFADLMQDFGGYEGQTIGS - - - - ASKLSKMMAAKARLLADVLEKALD - GYSDENNGAIDEASNTLYDQLKGFRDVG - - - -
 QFENLIKDFCTYIGQTI RS - - - - PKKLAEMMAGKARLLQNTLERALEQDIADDDNTE LNE - - - - - QYETFKDI - - - -
 DLIELFRGFFNHEAAPITN - - - - AKDFANALSAPTRYLKDAL - - - - - IQYQKDMQVSS - - - - - IFNNFKEY - - - -
 DLIELFRGFFNHEAAPITN - - - - AKDFANALSAPTRYLKDAL - - - - - IAYQKDDQVSS - - - - - IFKNFKEY - - - -
 DLIELFRGFFNHEAAPITN - - - - AKDFANALSAPTRYLKDAL - - - - - IAYQKDDQVSS - - - - - IFKNFKEY - - - -
 - - - - LADAMYNVMP TKLGD - RNYWENFTKKTGNIARTLNNRL - - - - - KIIFDKNPEFFH - - - - - G - - - - FLDSLREN - - - -
 DLIELFKSFFNHEAAPITN - - - - AKDFATHLSPR TKY LKDAL - - - - - IKYQEKAQVSS - - - - - IFNNFKEY - - - -
 DLIELFKSFFNHEAAPITN - - - - AKDFATHLSPR TKY LKDAL - - - - - IKYQEKAQVSS - - - - - IFNNFKEY - - - -
 DLNRLLVAFFDWQPAVIGEWRS AVQQFRVELPAI LGHLRERI - - - - - DKAYDDNEAFTAKATA - - - - - FLQHARET - - - -
 TLAPLIATFLQWSP IAPKS - - - - AKALAQVSARLCRLLRDEV - - - - - IE - QLELGSAG - LTE - - - - - LAKDWRHL - - - -
 ALRELFLDFLNWRPLVPRN - - - - PQELARFLAPLARFLREAV - - - - - LE - EVRENPN GELAR - - - - - LREEWRKN - - - -
 RVEELL LDFLHWQPLVPKN - - - - PQELARFLAPL TRFLREAV - - - - - VE - ALREDPEGRLAH - - - - - LYREWAGDPAS

	360	370	380	390	400	410	420	430
<i>Cchl</i>	FKATDVNSL	LLKDFRNATQQND	PIIHFYETFLAEYDPT	LRKSRGWWYTP	PEPVVNFIVRAVD	DDILK-TEF	-----	DLRDGLTDTSK
<i>FtnUV</i>	FRAVDLRKI	LSKFRSTKTQDP	IVHFDYEDFLSEYDS	KLRKAKGWWYTP	QPVVSFIVRAVDEV	LK-SEF	-----	GLSQGLADTTK
<i>Hpy300X</i>	INHVDMGSI	IKDL---NDDKDP	YLHFYETFLSAYDP	KLREKKGWYTP	PDSVVKFIINALD	SLLK-THFKDAP	PLGLKSALDN---	
<i>Hpy99XI</i>	INHVDMGPI	IKDL---NDDKDP	YLHFYETFLSAYDP	KLREKKGWYTP	PDSVVEFIINALD	SLLK-THFKDAP	PLGLKSALDN---	
<i>Hpy99XI</i>	INHVDMGPI	IKDL---NDDKDP	YLHFYETFLSAYDP	KLREKKGWYTP	PDSVVEFIINALD	SLLK-THFKDAP	PLGLKSALDN---	
<i>HpyAXV</i>	YESVKTAL	--HAKSQKSQQEL	IKNLYNTFFKEAF	KKQSEKLGIVYTP	IEVVDFILRATNG	ILK-KHF	-----	NTDFND---
<i>HpyAXV</i>	INHVDMDSI	LKDL---NDDKDP	YLHFYETFLSTYDP	KLRESKGVYTP	PDSVVKFIINALD	SLLK-THFKDAP	PLGLKSALDN---	
<i>HpyAXV</i>	INHVDMDSI	LKDL---NDDKDP	YLHFYETFLSTYDP	KLRESKGVYTP	PDSVVKFIINALD	SLLK-THFKDAP	PLGLKSALDN---	
<i>OgrI</i>	YQAIETAAA	--EITDHAEKQTF	LKVIYEGFYQSYNP	DAADRLGVVYTP	NEIVRFMVRATD	WLCE-RHF	-----	GKRLAD---
<i>RpaI</i>	LGEVNWVHI	-----SKDKPE	AWLYFYEDFL	LEVYDNTLRKKT	GSYYTPPEVVA	AMVRLADEALRG	ELF-----	GRPKGFAS---
<i>TaqII</i>	LRAVDPSVF	-----RVQGVDP	WLYFYEDFLQAY	DPDLRKDMGVYTP	PVPVVRAMVRL	VDEALK-EGF	-----	GLAEGLAH---
<i>TspGWI</i>	IAAVDPAHF	-----GGGGAD	PWLYFYEDFL	EAYDPELRKDM	GVYTPVPVVR	AMVQLVDDLLR	-TKM-----	GKPLGLAE---

Участок 2

	360	370	380	390	400	410	420	430
<i>Cchll</i>	FKATDVNSL	LKDFRNATQQNDP	I I H F Y E T F L A E Y D P T L R K S R G W W Y T P E P V V N F I V R A V D D I L K - T E F - - - - - D L R D G L T D T S K					
<i>FtnUV</i>	FRAVDLRKI	LSKFGRSTKTQDP	I V H F Y E D F L S E Y D S K L R K A K G W W Y T P Q P V V S F I V R A V D E V L K - S E F - - - - - G L S Q G L A D T T K					
<i>Hpy300</i>	I N H V D M G S I	I K D L - - - N D D K D P Y L H F Y E T F L S A Y D P K L R E K K G V Y Y T P D S V V K F I I N A L D S L L K - T H F K D A P L G L K S A L D N - - -						
<i>Hpy99XI</i>	I N H V D M G P I	I K D L - - - N D D K D P Y L H F Y E T F L S A Y D P K L R E K K G V Y Y T P D S V V E F I I N A L D S L L K - T H F K D A P L G L K S A L D N - - -						
<i>Hpy99XI</i>	I N H V D M G P I	I K D L - - - N D D K D P Y L H F Y E T F L S A Y D P K L R E K K G V Y Y T P D S V V E F I I N A L D S L L K - T H F K D A P L G L K S A L D N - - -						
<i>HpyAXV</i>	Y E S V K T E A L - -	H A K S Q K S Q Q E L I K N L Y N T F F K E A F K K Q S E K L G I V Y T P I E V V D F I L R A T N G I L K - K H F - - - - - N T D F N D - - -						
<i>HpyAXV</i>	I N H V D M D S I	L K D L - - - N D D K D P Y L H F Y E T F L S T Y D P K L R E S K G V Y Y T P D S V V K F I I N A L D S L L K - T H F K D A P L G L K S A L D N - - -						
<i>HpyAXV</i>	I N H V D M D S I	L K D L - - - N D D K D P Y L H F Y E T F L S T Y D P K L R E S K G V Y Y T P D S V V K F I I N A L D S L L K - T H F K D A P L G L K S A L D N - - -						
<i>Ogrl</i>	Y Q A I E T A A A - -	E I T D H A E K Q T F L K V I Y E G F Y Q S Y N P D A A D R L G V V Y T P N E I V R F M V R A T D W L C E - R H F - - - - - G K R L A D - - -						
<i>Rpal</i>	L G E V N W H V I - - - - -	S K D K P E A W L Y F Y E D F L E V Y D N T L R K K T G S Y Y T P P E V V A A M V R L A D E A L R G E L F - - - - - G R P K G F A S - - -						
<i>TaqII</i>	L R A V D P S V F - - - - -	R V Q G V D P W L Y F Y E D F L Q A Y D P D L R K D M G V Y Y T P V P V V R A M V R L V D E A L K - E G F - - - - - G L A E G L A H - - -						
<i>TspGWI</i>	I A A V D P A H F - - - - -	G G G G A D P W L Y F Y E D F L E A Y D P E L R K D M G V Y Y T P V P V V R A M V Q L V D D L L R - T K M - - - - - G K P L G L A E - - -						

	360	370	380	390	400	410	420	430
<i>Cchll</i>	FKATDVNSL	LKDFRNATQQNDP	I I H F Y E T F L A E Y D P T L R K S R G W W Y T P E P V V N F I V R A V D D I L K - T E F - - - - - D L R D G L T D T S K					
<i>FtnUV</i>	FRAVDLRKI	LSKFGRSTKTQDP	I V H F Y E D F L S E Y D S K L R K A K G W W Y T P Q P V V S F I V R A V D E V L K - S E F - - - - - G L S Q G L A D T T K					
<i>Hpy300</i>	I N H V D M G S I	I K D L - - - N D D K D P Y L H F Y E T F L S A Y D P K L R E K K G V Y Y T P D S V V K F I I N A L D S L L K - T H F K D A P L G L K S A L D N - - -						
<i>Hpy99XI</i>	I N H V D M G P I	I K D L - - - N D D K D P Y L H F Y E T F L S A Y D P K L R E K K G V Y Y T P D S V V E F I I N A L D S L L K - T H F K D A P L G L K S A L D N - - -						
<i>Hpy99XI</i>	I N H V D M G P I	I K D L - - - N D D K D P Y L H F Y E T F L S A Y D P K L R E K K G V Y Y T P D S V V E F I I N A L D S L L K - T H F K D A P L G L K S A L D N - - -						
<i>HpyAXV</i>	Y E S V K T E A L - -	H A K S Q K S Q Q E L I K N L Y N T F F K E A F K K Q S E K L G I V Y T P I E V V D F I L R A T N G I L K - K H F - - - - - N T D F N D - - -						
<i>HpyAXV</i>	I N H V D M D S I	L K D L - - - N D D K D P Y L H F Y E T F L S T Y D P K L R E S K G V Y Y T P D S V V K F I I N A L D S L L K - T H F K D A P L G L K S A L D N - - -						
<i>HpyAXV</i>	I N H V D M D S I	L K D L - - - N D D K D P Y L H F Y E T F L S T Y D P K L R E S K G V Y Y T P D S V V K F I I N A L D S L L K - T H F K D A P L G L K S A L D N - - -						
<i>Ogrl</i>	Y Q A I E T A A A - -	E I T D H A E K Q T F L K V I Y E G F Y Q S Y N P D A A D R L G V V Y T P N E I V R F M V R A T D W L C E - R H F - - - - - G K R L A D - - -						
<i>Rpal</i>	L G E V N W H V I - - - - -	S K D K P E A W L Y F Y E D F L E V Y D N T L R K K T G S Y Y T P P E V V A A M V R L A D E A L R G E L F - - - - - G R P K G F A S - - -						
<i>TaqII</i>	L R A V D P S V F - - - - -	R V Q G V D P W L Y F Y E D F L Q A Y D P D L R K D M G V Y Y T P V P V V R A M V R L V D E A L K - E G F - - - - - G L A E G L A H - - -						
<i>TspGWI</i>	I A A V D P A H F - - - - -	G G G G A D P W L Y F Y E D F L E A Y D P E L R K D M G V Y Y T P V P V V R A M V Q L V D D L L R - T K M - - - - - G K P L G L A E - - -						

	570	580	590	600	610	620	630	640
<i>CchII</i>	---HHPETGTL	-FASWLSQEANEANY	IKRDT	PVMVVLGNPPY	---	SGHSANKS	--KW	---
<i>FtnUV</i>	---HHPDTGTL	-FANWLSNEANEANQ	IKKDT	PVMVVMGNPPY	---	SGISSNTG	--EW	---
<i>Hpy300</i>	YE--IIAYRGLNPI	-FEKELSN---	AQEIKKNENIL	IITGNPPY	---	SGASENKGLFEWEVKATYG	---	IEPEFQTIEI
<i>Hpy99XI</i>	E--IIAYRGLNPI	-FEKELSN---	AQEIKKNENIL	IITGNPPY	---	SGASENKGLFEWEVKATYG	---	IEPEFQTIEI
<i>Hpy99XI</i>	E--IIAYRGLNPI	-FEKELSN---	AQEIKKNENIL	IITGNPPY	---	SGASENKGLFEWEVKATYG	---	IEPEFQTIEI
<i>HpyAXV</i>	--LEEKTNGVLP	LYEDL-KENKG	IKDTLANQN	IRVIIGNPPY	--SAGAK	SQNDNNQNL	---	SHPK---
<i>HpyAXVE</i>	--IIAYRGLSPI	-FEKELSN---	AQEIKKNENIL	IITGNPPY	---	SGASENKGLFEWEVKATYG	---	IDPKFQTIEI
<i>HpyAXVE</i>	--IIAYRGLSPI	-FEKELSN---	AQEIKKNENIL	IITGNPPY	---	SGASENKGLFEWEVKATYG	---	IDPKFQTIEI
<i>OgrI</i>	GLGIRRGQQMSFL	QGFTD---	ENTERVQAQNR	RKISVVIIGNPPY	--NANQQNENDNN	KNR	---	EYPR---
<i>Rpal</i>	---VEEESLGQV	--YEP	IAKSRRREANAV	VKKDKPITVV	IGNPPY	---	KEKAKGRG	--GWIESGSGDL
<i>TaqII</i>	---EAPPLEQVFF	YERLA	EERKRAAE	LKRDKPILVV	IGNPPY	DRVEGESQEERER	KG	--GWVLRGPREPY
<i>TspGWI</i>	---DAPPLEREFF	YERLA	QERREAA	RVKREVP	ILVILGNPPY	DRVEGESKEARK	ARG	--KWI
								VQGGKDPQDPNSPPP

Участок 3

```

      570      580      590      600      610      620      630      640
CchlI  - - - - HHPETGTL - FASWLSQEANEANY I KRDT PVMVVLGNPPY - - - - SGHSANKS - - KW - - - -
FtnUV  - - - - HHPDTGTL - FANWLSNEANEANQ I KKDT PVMVVMGNPPY - - - - SGISSNTG - - EW - - - -
Hpy300>E - - I I AYRGLNPI - FEKELSN - - - - AQE I KKNEN I L I I TGNPPY - - - - SGASENKGLFEWEVKATYG - - - - IEPEFQTI E I
Hpy99XI E - - I I AYRGLNPI - FEKELSN - - - - AQE I KKNEN I L I I TGNPPY - - - - SGASENKGLFEWEVKATYG - - - - IEPEFQTI E I
Hpy99XI E - - I I AYRGLNPI - FEKELSN - - - - AQE I KKNEN I L I I TGNPPY - - - - SGASENKGLFEWEVKATYG - - - - IEPEFQTI E I
HpyAXV - - LEEKTNKGVLP LYEDL - KENKGIKDTLANQNI RVI I GNPPY - - SAGAKSQNDNNQNL - - - - SHPK - - - -
HpyAXVE - - I I AYRGLSPI - FEKELSN - - - - AQE I KKNEN I L I I TGNPPY - - - - SGASENKGLFEWEVKATYG - - - - IDPKFQTI E I
HpyAXVE - - I I AYRGLSPI - FEKELSN - - - - AQE I KKNEN I L I I TGNPPY - - - - SGASENKGLFEWEVKATYG - - - - IDPKFQTI E I
OgrI   GLG I RRGQMSFLGQFTD - - - - ENTERVQAQNRRI SVVI GNPPY - - NANQQNENDNNKNR - - - - EYPR - - - -
Rpal   - - - - VEEESLGQV - - YEP I AKSRREANAVKKDKP I TVVI GNPPY - - - - KEKAKGRG - - GWI ESGSGGDL - - VAPM - - - -
TaqII  - - - - EAPPLEQVFFYERLAEERKRAAE LKRDKP I LVVLGNPPYDRVEGESQEERERKG - - GWVLRGPREPY - - - - PL - - - -
TspGWI - - - - DAPPLEREFFYERLAQERREAARVKREVP I LV I L GNPPYDRVEGESKEARKARG - - KWI VQGKKDPQDPNSPPP - - - -

```

```

      570      580      590      600      610      620      630      640
CchlI  - - - - HHPETGTL - FASWLSQEANEANY I KRDT PVMVVLGNPPY - - - - SGHSANKS - - KW - - - -
FtnUV  - - - - HHPDTGTL - FANWLSNEANEANQ I KKDT PVMVVMGNPPY - - - - SGISSNTG - - EW - - - -
Hpy300>E - - I I AYRGLNPI - FEKELSN - - - - AQE I KKNEN I L I I TGNPPY - - - - SGASENKGLFEWEVKATYG - - - - IEPEFQTI E
Hpy99XI E - - I I AYRGLNPI - FEKELSN - - - - AQE I KKNEN I L I I TGNPPY - - - - SGASENKGLFEWEVKATYG - - - - IEPEFQTI E
Hpy99XI E - - I I AYRGLNPI - FEKELSN - - - - AQE I KKNEN I L I I TGNPPY - - - - SGASENKGLFEWEVKATYG - - - - IEPEFQTI E
HpyAXV - - LEEKTNKGVLP LYEDL - KENKGIKDTLANQNI RVI I GNPPY - - SAGAKSQNDNNQNL - - - - SHPK - - - -
HpyAXVE - - I I AYRGLSPI - FEKELSN - - - - AQE I KKNEN I L I I TGNPPY - - - - SGASENKGLFEWEVKATYG - - - - IDPKFQTI E
HpyAXVE - - I I AYRGLSPI - FEKELSN - - - - AQE I KKNEN I L I I TGNPPY - - - - SGASENKGLFEWEVKATYG - - - - IDPKFQTI E
OgrI   GLG I RRGQMSFLGQFTD - - - - ENTERVQAQNRRI SVVI GNPPY - - NANQQNENDNNKNR - - - - EYPR - - - -
Rpal   - - - - VEEESLGQV - - YEP I AKSRREANAVKKDKP I TVVI GNPPY - - - - KEKAKGRG - - GWI ESGSGGDL - - VAPM - - - -
TaqII  - - - - EAPPLEQVFFYERLAEERKRAAE LKRDKP I LVVLGNPPYDRVEGESQEERERKG - - GWVLRGPREPY - - - - PL - - - -
TspGWI - - - - DAPPLEREFFYERLAQERREAARVKREVP I LV I L GNPPYDRVEGESKEARKARG - - KWI VQGKKDPQDPNSPPP - - - -

```

Как найти консервативные мотивы 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

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.

Задача 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

Задача 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.

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

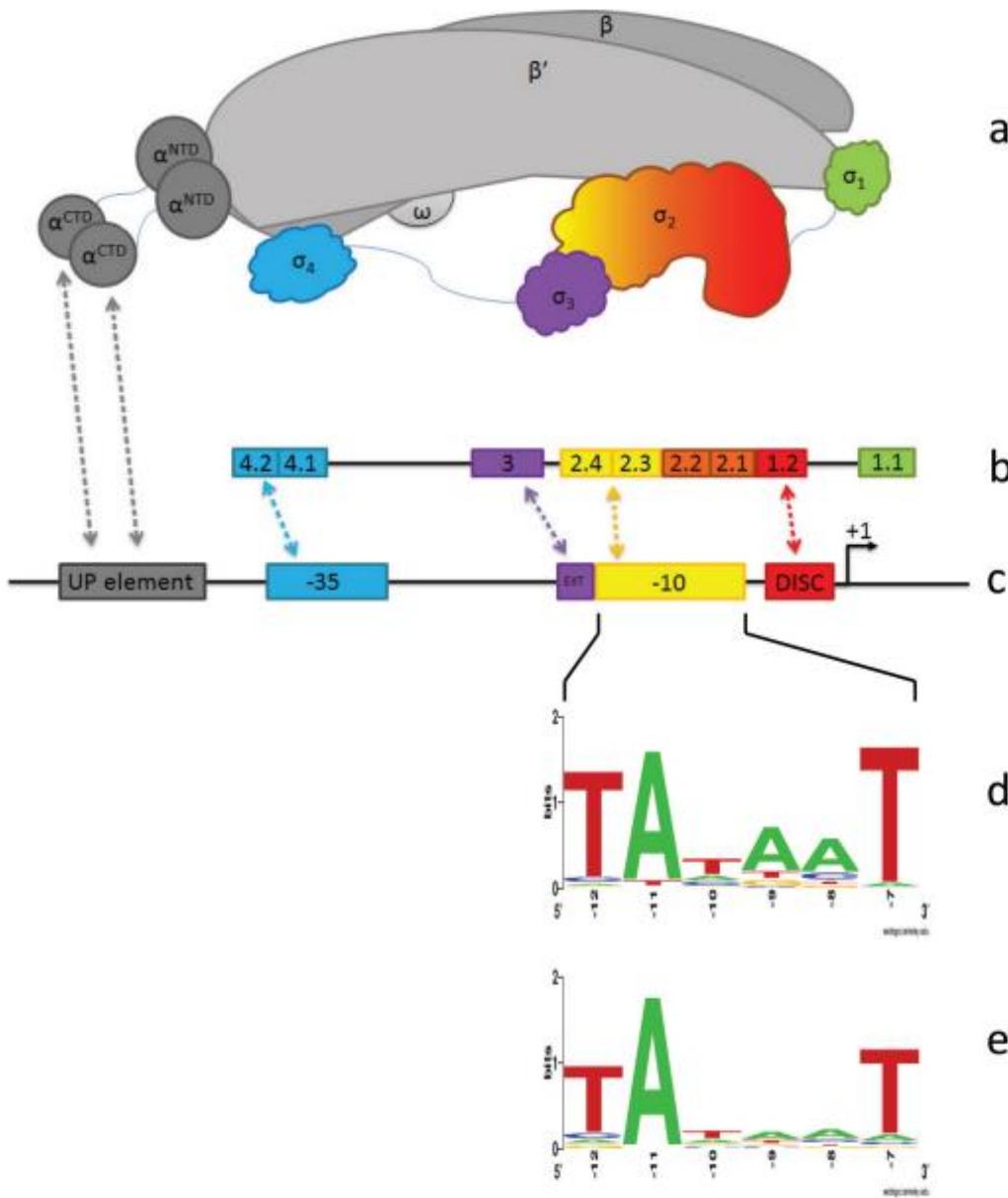


Fig. 1. Schematic representation of bacterial RNAP (RNA polymerase) holoenzyme, factor structure, and representative promoter structure. (a) Subunit architecture of holoenzyme (core subunits presented in various shades of grey); represented as 4 domains colour-coded according to (b) regions and subregions 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).

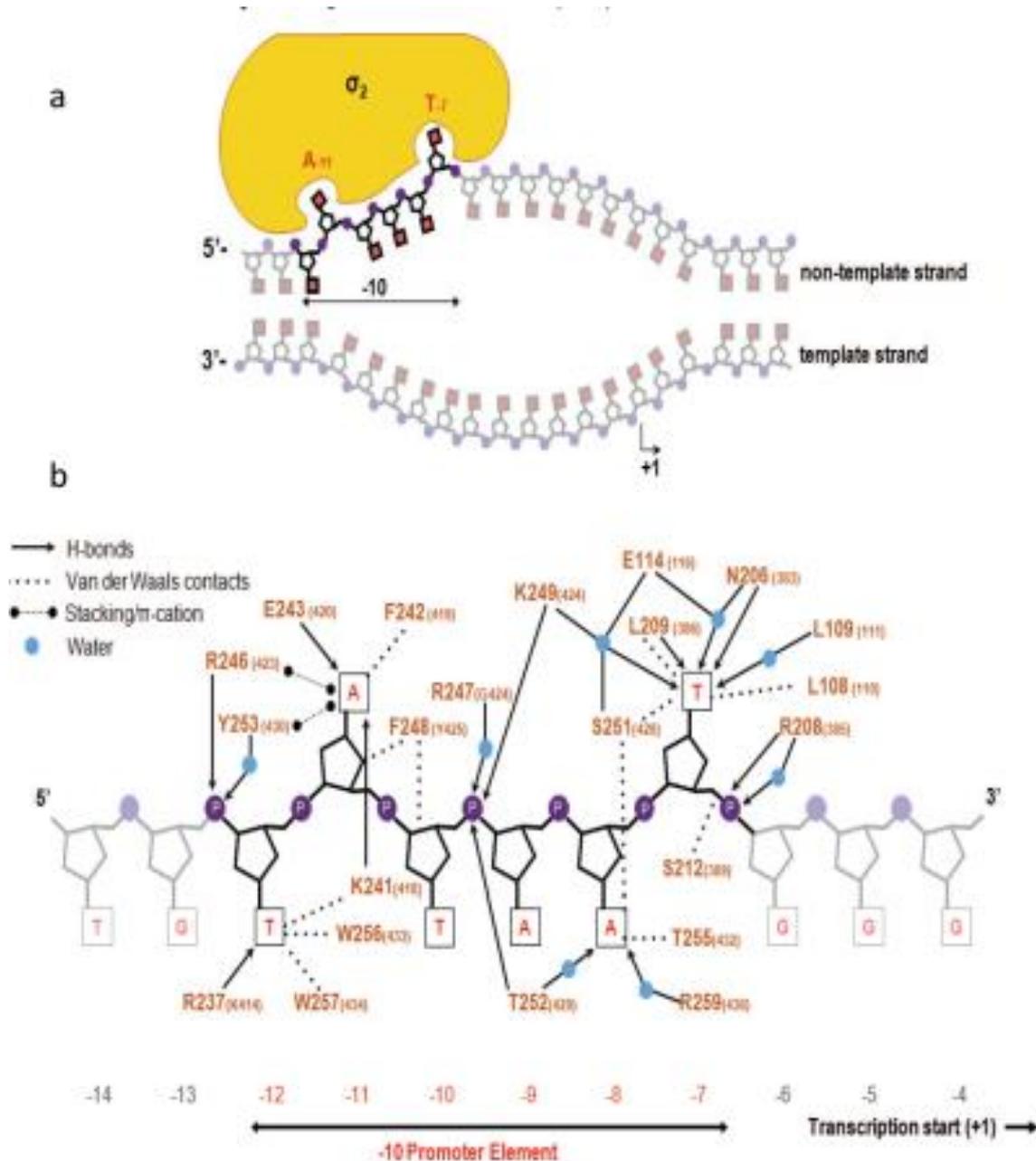


Fig. 3. Interaction patterns between *Thermus aquaticus* A 2 amino acid residues and -10 promoter element nucleotides. (a) Schematic diagram of open complex (represented in duplex DNA context) with 2 flipped out bases captured by domain 2. (b) Redrawing of inferred interactions between key amino acids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

IIIb. Паттерны

fuzzpro, fuzznuc, Jalview

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. Авторы – наши знакомые.
- Отстаёт по объёму коллекций от лидеров, но обгоняет их по аккуратности
- Для учебных целей (и не только) подходит

Банк ProSite. Сейчас недоступен.
Надолго ли – не знаю

<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 своей белке.
2. Искать свой мотив в коллекции последовательностей ProSite.
3. Искать свой мотив в своей 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

Ис. PSSM и PSI BLAST для белков

С.А.Спирин, А.С.Ершова

Вспомним PWM, вес и информационное содержание

TTATGCC
 ATCTTCA
 GTATTA

	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

выравнивание



PWM для данного
выравнивания

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

$$I = \sum_k I_k$$

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)

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

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

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:

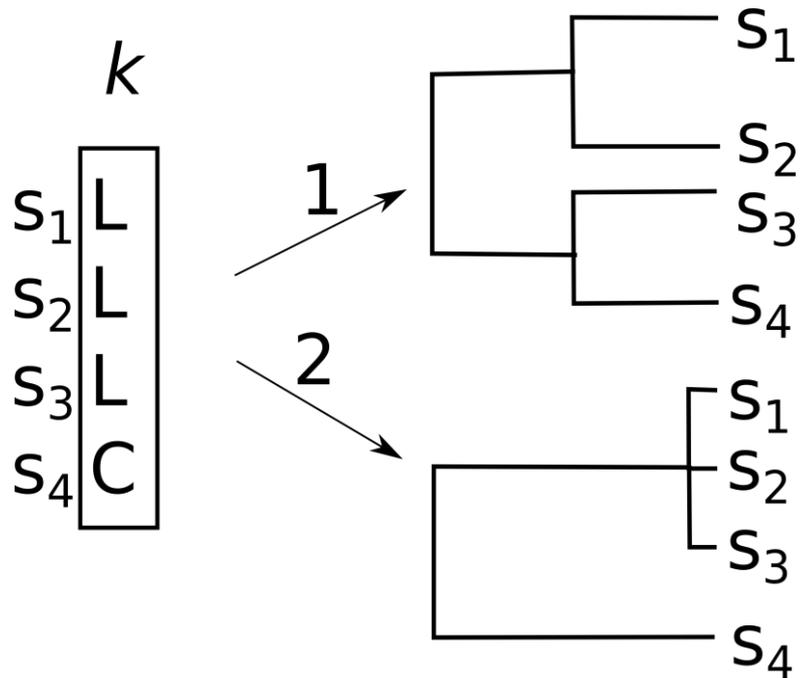
$$S_{ki} = \frac{1}{\lambda} \log \frac{f_{ki}}{p_i}$$

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:

$$f_{ki} = \frac{\sum_{s:a_{sk}=i} w_s + \psi_i}{\sum_s w_s + \sum_i \psi_i}$$

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 псевдоотсчетов 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

* gibbs sampling for motif finding

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

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

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☺)

$$P = \exp [(I(B) - I(A))/T]$$

Есть и другие алгоритмы

У А.А.Миронова с соавт
ChIPMunk (В.Макеев, И.Кулаковский с соавт.)
(<https://opera.autosome.ru/chipmunk/discovery>)

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

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

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_418484.4lyjcB	281	298	21.2367	5.3e-09	0.00758	AATTGTGATATAGTTCAC
1	NP_418485.1lyjcC	149	132	21.2367	5.3e-09	0.00758	AATTGTGATATAGTTCAC
1	NP_418031.1lyiaJ	175	158	19.8034	3.86e-08	0.0173	AAGTGTGCCGTAGTTCAC
1	NP_418032.1lyiaK	26	43	19.8034	3.86e-08	0.0173	AAGTGTGCCGTAGTTCAC
1	NP_418535.1lproP	37	54	19.7078	4.3e-08	0.0173	ATGTGTGAAGTTGATCAC
1	NP_414666.1lgcd	126	143	19.6123	4.85e-08	0.0173	AATTGTGATGACGATCAC
1	NP_414667.4lhpt	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.

Метод «по бернулли» не всегда работает

- В задании 6.4, те, кто его выполнял, вероятно, определяли ожидаемое число слов, (например, слова СТАГ) в геноме, путем перемножения частот отдельных букв.
- Такой способ подразумевает, что геном устроен как бернуллиевская случайная последовательность: появление следующей буквы не зависит от предыдущей.
- Это предположение нарушается.
- Пример. Число слово ТА в большинстве геномов меньше, чем ожидаемое по бернулли. Поэтому число слов СТАГ в геномах будет менее ожидаемого по бернулли!
- Как учесть этот эффект – учитывать частоты подслов!

GATC
CCNGG
GGWCC

Compositional Bias (CB) паттерна =
(наблюдаемое число)/ожидаемое
(называют также контрастом)

CB >> 1 - слово перепредставлено

CB << 1 - слово недопредставлено

Вопрос как правильно вычислять «ожидаемое»

ММ. Марковская цепь максимального порядка

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

F_o – observed (наблюдаемое
число слов)

F_{exp} – expected (ожидаемое
число слов)

$$F_{exp}(w_1 \dots w_m) = F_o(w_1 \dots w_{m-1}) F_o(w_2 \dots w_m) / F_o(w_2 \dots w_{m-1}).$$

Можно объяснить так (для СТАГ). Перед словом ТА с частотой $F_o(СТА)/F_o(ТА)$ стоит буква С, с частотой $F_o(ТАГ)/F_o(ТА)$ после ТА стоит буква Г. В предположении независимости имеем:

$$F_o(СТАГ) = F_o(СТА)/F_o(ТА) * F_o(ТА) * F_o(ТАГ) / F_o(ТА) = \\ = (F_o(СТА) * F_o(ТАГ)) / F_o(ТА)$$

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

К. Метод Карлина с соавт.

Учёт частот всех подслов, включая разрывные

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

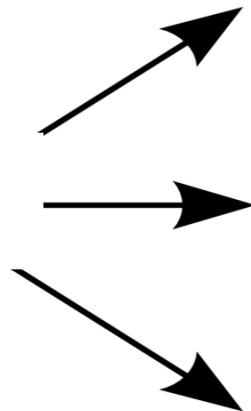
$$F_{exp} = \frac{\prod F_o(\text{odd } N)}{\prod F_o(\text{even } N)}$$

GATN, GANC, 1N
GNTC, NATC

GANN, GNNC, NNTC, 2N
NATN, NANC, GNTN

GNNN, NANN, 3N
NNTN, NNNC

CB(GATC) =



КОНЕЦ ПРЕЗЕНТАЦИИ