**Результаты проверки теста по написанию текста.**

1. Тема: HMM профиль выравнивания
2. Тема: информационное содержание выравнивания
3. Тема: позиционная весовая матрица
4. Тема: сравнение PSSM и HMM профилей
5. Тема: база данных Pfam
6. Тема: сигналы и мотивы
7. Тема: поиск сигналов в ДНК

**Условные обозначения**

Фраза нет замечаний

Слово,словосочетание, фраза неудачное, лучше подобрать более точное; или неправильное по смыслу

Вставка нужна или замечание

Слово убрать, лишнее

Фраза убрать, не нужна

║ разделить фразу на две

 Переставить слово, словосочетание или фразу

[две мысли] мои замечания и предложения написаны красным шрифтом и в скобках [ ]

**5/5** оценка за содержание/за текст

Число слов – последнее поле в 1й строке. Превышение на одно-два слова прощаю, т.к. может объясняться особенностями алгоритма выделения слов в Word.

X====================================================X

**5/5** 10/04/2020 09:19:46 **russal2010** **2 IC** 120

По выравниванию участков генома (или белка) бывает нужно понять, может ли оно иметь какое-то значение. Удобным способом предсказания является информационное содержание. Оно основывается на оценке того, насколько «необычно» видеть данные элементы выравнивания (буквы) на конкретных позициях. Если окажется достаточно необычно (распределение не случайное [распределение чего?]), есть основание предполагать, что в выравнивании содержится что-то важное. (Это не значит, что для любого важного сигнала информационное содержание будет большим: сигналы могут быть закодированы не только в виде консервативных участков последовательности). В качестве критерия «необычности» принято брать логарифм отношения наблюдаемой частоты данной буквы на этой позиции выравнивания к её средней частоте в геноме, причём для возможности учёта всех букв на позиции выравнивания «необычность» каждой берётся с весом, равным наблюдаемой частоте буквы на данной позиции.

**5/5** 10/04/2020 09:20:32 **zubareva.valeria 3 PWM** 121

Если нам дано выравнивание без гэпов, то для него мы можем построить позиционную весовую матрицу. Такая матрица позволяет оценить, частоту конкретной буквы на конкретной позиции относительно общей частоты каждого нуклеотида в геноме ║и используется для поиска сигналов [две мысли]. Матрица строится по следующему алгоритму:

1. считается общее число букв в каждой колонке выравнивания

2. считается частота каждого нуклеотида в данной позиции

3. добавление псевдокаунтов

4. берется натуральный логарифм отношения наблюдаемой частоты к ожидаемой

Для борьбы с нулевыми частотами используются псевдокаунты. Добавим каждому нуклеотиду в данной позиции очень малую частоту по сравнению с количеством букв (частота подбирается опытным путем). Мы имеем на это право так как, если количество последовательностей мало, то последовательности с другим нуклеотидом в данной позиции могли просто не попасть в выравнивание.

**5/5 10/04/2020 09:24:27 grib**.irina **6 сигнал** 120 Как рибосома узнает мРНК? Малая субъединица узнает кэп и сканирует РНК до старт-кодона.

У млекопитающих это обычно ATG, но, например, у SARS-CoV-2 первый ген начинается с 266 bp, а ATG есть и на позиции 107, и 266. Но окружение ATG на позиции 266 больше похоже на последовательность Козак - сигнал начала транскрипции у эукариот.

А какие типы сигналов вообще могут быть? Закодированные последовательностью (нуклеотидов или аминокислот), структурой НК или белка (G-квадруплекс, IRES – особым образом уложенная шпилька РНК), химической модификацией НК или белка (кэп, метилирование ДНК)

Мотив - описание последовательностей одного сигнала. Мотивом может быть: точная последовательность (AAAAAAAAAAAAAAAAAAААА (от десятков до сотен букв), CpG, GATC), паттерн (CCNGG, Козак для вида), выравнивание (консенсус, LOGO, позиционная весовая матрица (PWM), профиль (чаще для белков))

**5/5** 10/04/2020 09:24:32 **shilmaria 6 сигнал** 93 Сигнал – это «знак», который может быть закодирован последовательностью биополимера (белка или нуклеиновой кислоты), содержаться в различных их химических модификациях (например, метилирование ДНК) или же быть закодирован структурой биополимера (G-квадруплекс). Сигнал несет информационную нагрузку и считывается молекулами в клетке с биополимера ║ (например, промотор – последовательность ДНК, узнаваемая РНК-полимеразой, необходимая для начала транскрипции).

Так как нуклеиновые кислоты подвержены мутациям, а вслед за ними изменяются и белки, было бы даже странно, если бы все огромное многообразие сигналов в клетке было представлено лишь одним вариантом. Так и не происходит. Мотив – это совокупность возможным вариантов для сигнала в клетке.

**5/5** 10/04/2020 09:25:59 **esiling 2 IC 118**

 Согласно Вашей презентации, информационное содержание (IC) выравнивания отражает то, насколько данное выравнивание отличается от выравнивания случайных фрагментов той же длины и в том же числе. Это понятие определяют также как матожидание веса в колонке при распределении вероятностей букв b, заданных частотами букв в колонке, где f(b,j) – частоты, а по сути вероятности, букв на позициях, w(b,j) – вес из матрицы PWM без псевдоотсчетов. Так, IC – это произведение частоты на вес, и это всегда неотрицательное число. Чтобы найти IC, надо сначала посчитать число букв на позицию, а потом определить их частоты без псевдокаунтов, затем подставить f и w в формулу.

Большое значение IC говорит о том, что частоты букв в колонке сильно отличаются от случайных, и информации в этой колонке много.

**4/4** 10/04/2020 09:26:00 **mariasha.k.1012 3 PWM 121**

Позиционная весовая матрица (PWM) – это матрица, которая строится по готовым, выровненным без гэпов последовательностям и используется для поиска последовательностей в геноме. Чтобы построить эту матрицу, сначала необходимо посчитать количества каждой буквы в каждом столбце выравнивания, затем посчитать частоту каждой буквы в колонке, поделив полученные количества каждой буквы в колонке на общее число букв в колонке (или на количество последовательностей, что то же самое). Чтобы в итоге посчитать вес каждой буквы, надо взять полученную частоту для каждой буквы и частоту встречаемости этого нуклеотида в геноме, тогда натуральный логарифм от отношения частоты буквы к частоте нуклеотида в геноме будет весом данной буквы в колонке. Используя эти веса, можно построить консесуную последовательность и паттерн для поиска похожих последовательностей в геноме (или любой другой последовательности).

**5/4** 10/04/2020 09:26:49 **elizaveta-biloba 3 PWM 120**

Позиционная весовая матрица - способ описания качества выравнивания, позволяющий грубо оценить силу мотива. Для нее используется выравнивание без гэпов относительно коротких фрагментов нуклеиновой кислоты, в которых содержится наш мотив. Для каждой позиции для каждой буквы высчитывается вес, который учитывает базовые частоты оснований в геноме и консервативность позиции выравнивания. Сначала считается абсолютное количество букв в позиции. Так как в расчете веса используется логарифмирование, есть смысл добавить абсолютную погрешность к этим величинам (в таком случае считаем псевдокаунты). Для каждой позиции рассчитывается частота каждой буквы. Высокая частота не всегда обоснована важностью и консервативностью позиции, поэтому она нормализуется на базовые частоты оснований (деление), ║ и далее это отношение логарифмируется для удобства работы. Для любой последовательности можно определить похожесть на профиль исходя из весов в матрице.

**5/4.5** 10/04/2020 09:26:53 **annapago 1 HMM 120**

Профиль HMM – это описание выравнивания. Программа, создающая его, строит по выравниванию автомат по построению последовательностей. Для каждой буквы или гэпа считается вероятность быть в данной позиции, ║ также подсчитывается вероятность перехода из одной позиции в другую, ║ при этом разрешены индели, то есть длина не фиксирована. Потом вместо вероятностей подсчитываются веса: эмиссионные (для позиции) делятся на частоту буквы и от получившегося значения берется [натуральный] логарифм. От вероятностей перехода тоже берётся натуральный логарифм. Теперь эта система умеет рассчитывать вес для конкретной последователности суммированием веса по позициям и переходам. После этого профиль калибруется по случайному банку для нормализиции веса и расчёта e-value. Таким образом по профилю можно оценить, насколько данная последовательность похожа на выравнивание, что может использоваться для поиска гомологичных доменов белков по базам данных.

**5/5** 10/04/2020 09:27:37 **kawansit1 6 сигнал** 98 В последовательностях нуклеиновых кислот и белков встречаются многочисленные короткие фрагменты, имеющие некую биологическую роль, в связи с чем, слабо меняющиеся в ходе эволюции. Их объединяют термином «мотивы». Мотивы (особенно в белках) могут быть и структурными, то есть консервативной является не столько последовательность, сколько структура определенного фрагмента молекулы. Такой мотив может отвечать за связывание ДНК или иона в ионном канале. Под сигналами понимают консервативные последовательности, адресованные неким эффекторам, распознающим их и запускающим определенные процессы. Сигналом может быть участок, с которым связываются факторы, запускающие транскрипцию ДНК или сигнальный пептид в белке, отвечающая за транспорт этого белка в определенный органоид клетки.

**5/5** 10/04/2020 09:28:00 crasyempress3090 **5 Pfam 121**

База данных Pfam - база данных семейств белковых доменов. Каждое из семейств представлено множественным выравниванием фрагментов белковых последовательностей и HMM. Основана база данных в 1997 году в Институте Сенгера. Записи в Pfam подразделяются на шесть типов: Family - набор гомологичных участков, Domain - функциональный участок, встречающийся в различных белковых архитектурах, Repeat - короткий участок, который является неустойчивым в изоляции, но образует стабильную структуру, когда присутствуют несколько его копий, Motif - короткий консервативный участок вне глобулярных доменов, Coiled-Coil - области, формирующие суперспирали, то есть пучки из 2-7 скрученных альфа-спиралей; Disordered - консервативные участки со смещенным аминокислотным составом, не формирующие устойчивой (глобулярной) структуры.

Что можно делать в Pfam: можно узнать доменную архитектуру последовательности белка, можно найти белок по доменной архитектуре и многое другое.

**4/5** 10/04/2020 09:28:09 **olga.shigal 6 сигнал** 101 Сигналы – это последовательности ДНК, которые читаются белками. Это тем или иным способом приводит к изменению процессов, в которых участвует данная молекула ДНК. Сигналы могут иметь разную силу, т.е. вероятность того, что при наличии данного мотива, или сигнала, произойдет то, что мы ожидаем. Чем сильнее сигнал, тем меньше вероятность того, что он появится в геноме случайно. Для того, чтобы оценить силу сигнала, можно посчитать его информационное содержание. Кроме того, можно оценивать представленность определенных паттернов, мотивов, вычисляя отношение числа их ожидаемых вхождений в геном к числу наблюдаемым. Соответственно, чем выше отклонение числа наблюдаемых вхождений, тем более вероятно, что данный мотив выполняет какую-то функцию.

**4/5** 10/04/2020 09:28:33 **yakushev.alex 6 сигнал** 72 Сигнал - это фрагмент макромолекулы, несущий информацию о каком-либо свойстве этой макромолекулы или ее участка для других белков. Примеры сигналов - последовательность Шайна-Дальгарно и сигнал ядерной локализации.

Мотив является более общим понятием.[бывают сигналы, которые не определяются мотивом. Например, шпилька РНК] Это любая последовательность, слабо меняющаяся в процессе эволюции. Сигналы также являются мотивами, но не описывают их множество полностью. Например цинковые пальцы в ДНК-распознающих белках являются мотивами, но сигнала об их свойстве для других макромолекул, по всей видимости, не несут.

**5/5** 10/04/2020 09:28:41 **mtochilkina 1 HMM** 59

HMM профиль выравнивания – это описание выравнивания. В какой-то степени профиль похож на PWM для нуклеотидных последовательностей, но есть некоторые различия: в выравнивании для HMM профиля допускаются индели, алгоритм построения профиля другой. Профиль строится на основе seed-выравнивания(с инделями), далее калибруется по случайному банку для нормализации веса и расчёта E-value. HMM профиль используют для поиска доменов и доменных архитектур в белках.

**5/5** 10/04/2020 09:29:10 **vera.sdrv 1** 75

HMM (Hidden Markov Model) профиль - это описание множественного выравнивания с инделями (локального или глобального), с помощью которого можно осуществлять поиск доменов в белковых последовательностях. Для этого необходимо откалибровать профиль по банку случайных последовательностей для нормализации веса и e-value. Если про какие-то последовательности точно известно, что они содержат домен, то HMM профиль поможет уточнить порог нормализованного веса для находки - последовательности С доменом. Информация заимствована в вольной форме из презентации и рассказа А. В. Алексеевского.

**5/4** 10/04/2020 09:29:25 **tirena 5 Pfam** 119

База данных Pfam содержит в себе семейства белковых доменов, определённые с помощью MSA (Multiple Sequence Alignment) и HMM (Hidden Markov Model). Работая с Pfam, можно выполнять следующие задачи:

1) Найти домены Pfam в введённой белковой последовательности и отобразить её доменную структуру.

2) По ID или AC найти информацию (выравнивания предполагаемых гомологичных белков, распределение по таксонам) о семействе или клане доменов (объединении нескольких семейств по признаку общности происхождения).

3) Найти белки со специфической доменной структурой.

Кроме этого, разметка доменов Pfam полезна при совместном использовании с другими базами данных: в Uniprot и аналогичных базах белковых последовательностей можно задать поиск и выдачу белков содержащих или не содержащих определённые домены Pfam, а при дополнительном использовании PDB размеченные домены можно визуализировать на 3D структуре."

**4/3** 10/04/2020 09:29:29 fp.delta **5 Pfam** 69

Семейный класс по умолчанию, который просто указывает на то, что члены связаны между собой. Домены определяются в качестве автономного структурного блока или блока многократного использования последовательности, которые могут быть найдены в различных окружениях белка. Повторы обычно не являются стабильными в изоляции, а, как правило, требуется для образования тандемных повторов, с тем чтобы сформировать домен или расширенную структуру. Мотивы, как правило, более короткие блоки последовательности, найденные за пределами глобулярных доменов.

**5/3** 10/04/2020 09:29:30 **sonya-man 1 HMM** 81

Сначала скажу, что такое профиль. Профиль- описание выравнивания, которое применяется для поиска доменов в последовательностях белков. Как создать HMM(Hidden Markov Model) профиль выравнивания? 1) Создаем автомат для генерации последовательностей по нужному выравниванию; 2) Нужно определить вероятность сгенерировать последовательность автоматом для каждой из входных последовательностей и если она(вероятность) будет больше порогового значения, то последовательность соответствует профилю. Под словом «автомат» подразумеваем инструмент, который может создавать последовательности(случайные) определенной длины(конечной),с большой вероятностью схожие с выравниванием. В дз мы используем для построения HMM профиля пакет HMMER.

**5/5** 10/04/2020 09:29:30 **ivanmike 1 HMM** 102

 "HMM профиль - описание выравнивания, используемое для поиска доменов. Профиль схож с PWM, однако, в последовательностях в выравнивании возможны индели. В презентации signals-aal-2020s\_4 представлен следующий алгоритм построения профиля и поиска доменов:

Для построения профиля необходимы выравнивания seed, составляемые из сходных фрагментов разных белков. После построения с помощью специальной программы (например, пакет HMMER), профиль калибруется для нормализации расчета E-value и веса. Если есть множество последовательностей, для которых известно, есть ли в них домен или нет, можно рассчитать порог нормализованного веса для находки. Затем в базе данных последовательностей (например, Uniprot) с помощью HMM профиля находятся участки с весом больше порога, т.е белки, содержащие домен."

**5/5** 10/04/2020 09:29:38 elena-pt **5 Pfam** 121

 База данных Pfam - база семейств белковых доменов , ║ состоящая из нескольких частей. В Pfam A содержатся семейства, курируемые вручную. В Pfam B - семейства, генерируемые автоматически. Если нас интересует какой-то домен и архитектуры, в которые он включен, мы можем ввести в окно поиска его идентификатор, а затем выбрать вкладку Доменные архитектуры. Тогда мы увидим графическое представление всех архитектур, содержащих домен, вместе с данными о количестве последовательностей, которые их содержат. То же можно узнать про отдельный домен. Кроме того, можно посмотреть: 1) домены того же клана (клан — объединение семейств, имеющих общее эволюционное происхождение), 2) LOGO для данного домена (логотип последовательности, построенный на основании выравнивания: в каждой колонке графика мы видим несколько букв, высоты которых зависят от их информационного содержания каждой позиции).

**5/5** 10/04/2020 09:29:49 adodonova 6 сигнал 99 Сигналы – что-то особенное (последовательность нуклеотидов, аминокислот, структура, химическое строение), отличающее данное место ДНК, РНК или белка от соседей и имеющее биологическую функцию. Оно может привлекать ферменты, как в случае с бактериальным промотором (сигма-субъединица), кэпом (полимераза) или метилированными нуклеотидами (MutH). Также оно [оно это особенное? Ха-ха]может нести самостоятельную функцию, как ро-независимая терминация трансляции, IRES (кэп-независимая инициация трансляции) или цинковые пальцы (взаимодействие с ДНК и РНК). Если сигнал – последовательность нуклеотидов, то мотив – описание этой последовательности. Он может быть точным (поли-а-хвост), каким-то легко распознаваемым паттерном (сайты связывания эндонуклеаз рестрикции) или выравниванием в более сложном случае, и тогда представляется в виде консенсуса, LOGO, PWM или профиля.

**\_\_\_** 10/04/2020 09:29:49 **sobbolik 3 PWM** 113 СПИСАНО У sobbolik! :)))

 "PWM – позиционная весовая матрица. Она строится для поиска мотивов, то есть высоко консервативных последовательностей, регулирующих работу клетки. PWM строится по множественному выравниванию последовательностей. Причем последовательности в выравнивании должны быть получены из близкородственных организмов или из участков генома, выполняющих одинаковые функции.

Как строится матрица:

1) По выравниванию необходимо посчитать, сколько раз встретился каждый нуклеотид в каждой позиции. (N(b,j))

2) Считаем, как часто встречается на данной позиции каждая буква. Для этого делим N(b,j) на число последовательностей в выравнивании. (f(b,j)=N(b,j)/N)

3) В связи с тем, что повышенная частота буквы может объясняться повышенной частотой в геноме, мы производим следующие вычисления - w(b,j)=ln(f(b,j)/частоту b в геноме

Готово. К сожалению, не хватает слов написать, как PWM (как мотив-то найти)

**5/5** 10/04/2020 09:29:52 **tatoshik312 2 IC** 115

 Чтобы понять, насколько данное выравнивание отличается от случайного данной длины и такого же количества, используют информационное содержание (IC). Чтобы его подсчитать, строим таблицу, где столбцы – номер столбца в выравнивании, строки – один из нуклеотидов, считаем количество каждого нуклеотида в каждом столбце выравнивания и в соответствующую ячейку записываем частное ((f(b,j)) от деления встречаемости данного нуклеотида в столбце к числу всех последовательностей в выравнивании. Как написано в презентации Алексеевского А. В., величина IC для нуклеотида b в столбце j составляет IC(b,j)=f(b,j)\*log2[f(b,j)/p(b)]=f(b,j)\*w(b,j). Отмечу, что если ожидаемая частота (p(b)) равна наблюдаемой или наблюдаемая равна 0, то величина IC обращается в нуль. Получаем IC(j), складывая все значение IC в столбце. Оцениваем значимость этого столбца по тому, насколько получилось большим значение IC(j).

**5/5** 10/04/2020 09:30:07 **denwer02** **5 Pfam** 119

"Pfam — база данных для белковых доменов. Для каждого семейства в базе имеется множественное выравниванием фрагментов белков и HMM - профиль. Семейства делят на два типа: Pfam - A и Pfam - B. Также, особенностью Pfam является то, что в базе нет белков, имеющих хотя бы одну аминокислоту, принадлежащую к двум разным доменам. Семейства, имеющие общее эволюционное происхождение и сходные структурно, объединяют в кланы.

Pfam-A курируется вручную, для семейства строится seed - выравнивание из наиболее репрезентативных последовательностей, по нему создается HMM - профиль, далее производится поиск по Pfamseq и последовательности, у которых показатель соответствия HMM – профиль [профилю, что ли?] выше порогового значения, добавляются в полное выравнивание. Для большинства семейств есть функциональная аннотация.

Семейства Pfam-B генерируются автоматически с использованием алгоритма кластеризации доменов ADDA."

**5/5** 10/04/2020 09:31:18 sofyagdk26 2 IC 104

Когда строится выравнивание, утверждается, что последовательности гомологичны, то есть имеют общего предка. Хочется понимать, насколько это утверждение биологически обосновано для данного выравнивания. Для этого вычисляется «Информационное содержание» - оценка того, насколько невероятно получить такое выравнивание случайно, то есть выбрав рандомные участки (из базы данных/из генома).

Для подсчета IC буквы мы ее частоту в колонке делим на ожидаемую частоту. За ожидаемую частоту чаще всего берется ожидаемая частота в геноме – действительно, если бы мы выравнивали случайные последовательности из генома, то частота в колонке не отличалась бы от частоты буквы в обычной последовательности.

Далее от данного нормированного числа берется двоичный логарифм и умножается на частоту данной буквы.

Чтобы найти IC выравнивания нужно сложить IC для всех букв в колонке, потом сложить IC колонок.

**5/5** 10/04/2020 09:32:09 **ivanpodd 3 PWM** 77

Позиционная весовая матрица (PWM) – таблица, которая строится по некоторому выравниванию нескольких нуклеотидных последовательностей без гэпов. Такая матрица учитывает частоту каждой буквы в позиции и относит ее к частоте такой буквы в геноме. Логарифм этого отношения помогает решить, является ли появление такой буквы в данной позиции случайностью или же она имеет какое-то биологическое значение. Так, например, с помощью PWM можно найти повторяющиеся в геноме сигналы и определить их косенсусную последовательность. Построение такой матрицы используется в работе программы MEME.

**5/5** 10/04/2020 09:32:21 **sobbolik 3 PWM** 114

PWM – позиционная весовая матрица. Она строится для поиска мотивов, то есть высоко консервативных последовательностей, регулирующих работу клетки. PWM строится по множественному выравниванию последовательностей. Причем последовательности в выравнивании должны быть получены из близкородственных организмов или из участков генома, выполняющих одинаковые функции.

Как строится матрица:

1) По выравниванию необходимо посчитать, сколько раз встретился каждый нуклеотид в каждой позиции. (N(b,j))

2) Считаем, как часто встречается на данной позиции каждая буква. Для этого делим N(b,j) на число последовательностей в выравнивании. (f(b,j)=N(b,j)/N)

3) В связи с тем, что повышенная частота буквы может объясняться повышенной частотой в геноме, мы производим следующие вычисления - w(b,j)=ln(f(b,j)/частоту b в геноме

Готово. К сожалению, не хватает слов написать, как интерпретировать PWM (как мотив-то найти)"

10/04/2020 09:33:14 **ar-nikolya 3 PWM** 124

**5/4** "PWM(позиционная весовая матрица) - матрица, столбцы которой соответствуют позиции в последовательности, а строчки - буквам в алфавите. Значениями этой матрицы являются частоты встречаемости данной буквы на данной позиции.[нет] PWM строят на основе множественного выравнивания последовательней ║ с целью обнаружить новый мотив.[где?]

PWM строится в несколько шагов: 1) Сначала нужно посчитать, сколько раз каждая буква встречается на данной позиции в выравнивании - N(i,j). 2) Вычисляется частота встречаемости данной буквы на данной позиции. На этом этапе часто прибегают с псевдоотчетам, чтобы избавится от нулевой частоты буквы. Псевдоотчет - малое число относительно N [N не определено] и N(i,j) 3) На основе GC состава генома организма находим логарифм отношения правдоподобия как вес различия наблюдаемой частоты и ожидаемой. Это и есть матрица PWM, которая содержит оценку вероятности встречи данной буквы в данной последовательности."

**5/5** 10/04/2020 09:33:30 **batyrsha 3 PWM** 113

 "PWM это прежде всего числовая таблица.

Зачем она нужна?

Допустим, есть выравнивание последовательностей сигнала. Мы хотим найти все сигналы в интересующем нас геноме. Для этого нужно знать, похожа ли новая последоватнльность на исходный сигнал. Также PWM, (по множественным выравниваниям) используются для описания семейств белков.

Как её строить?

1. Берем выравнивание.

2. Считаем число букв в позиции

3. Считаем частоты букв (кол-во каждой буквы в столбце делим на общее кол-во букв столбца)

4. Считаем логарифм отношения правдоподобия

w(буква, позиция) = ln(наблюдаемая частота/ожидаемая)

5. В конце концов получаем матрицу весов PWM, где некоторые позиции получились inf

С ними мы боремся с поиощью псевдокаунтов. eps(b) должны быть оч маленькими по отношению к N; подбираются опытным путем

**5/5** 10/04/2020 09:33:40 **dr.alexey7 1** 76

HMM профиль – это конструкция, которая строится по выравниванию, которое может содержать последовательности разной длины. Конструкция состоит из «генератора» последовательностей, похожих на заданное выравнивание. «Генератор» создается следующим образом: для каждой позиции в выравнивании оценивается вероятность появления на этом месте каждой из букв (по их частотам) и вероятность гэпа . Также если у нас есть некоторая последовательность, мы можем оценить вероятность, с которой она могла бы быть сгенерирована, и если эта вероятность существенная, то последовательность считается соответствующей профилю.

**5/4** 10/04/2020 09:33:57 **nikit00000s 3 PWM** 82

 Это [кто?] такая матрица, которая помогает в поиске мотивов (описания последовательностей сигнала - точная последовательность, паттерн или выравнивание) в биологических последовательностях. ПВМ [везде либо ПВМ, либо PWM] строится на множественном выравнивании чем-то схожих последовательностей. Строится следующим способом: для каждой колонки выравнивания считается количество разных букв алфавита, затем вычисляется двоичный логарифм отношения наблюдаемого количества данной буквы к ожидаемому количеству буквы (в данной позиции). Строка->буква, колонка->позиция. В более сложном случае используются псевдоотчеты - они позволяют избежать проблемы с бесконечностью логарифма в нуле. Также на основе PWM можно выявить информационное содержание.

**4/4** 10/04/2020 09:34:48 **spyro 6 сигнал** 44

 "Сигнальные последовательности нужны для выполнения биологических процессов.

Примеры сигналов: 5'—кэп и 3'-поли-А конец — сигналы мРНК для старта транскрипции, [трансляции]║которые узнает рибосома.

Мотивы — короткие последовательности нуклеотидов/аминокислот, которые в процессе эволюции либо меняются слабо, либо не меняются вовсе. Предполагается, что они имеют в молекуле какую-то функцию."

**5/4** 10/04/2020 09:35:24 **d.nogina 5 Pfam** 44

База семейств белковых доменов [нет сказуемого]. Есть курируемая (A) и автоматическая части (B). + Кланы - объединения семейств белковых доменов, имеющих общее эволюционное происхождение. В HMM по поиску последовательности можно выяснить доменную структуру. 6 типов записей, в том числе домен, повтор, мотив, неструктурированный блок и т.п. [нет сказуемого]

**5/3** 10/04/2020 09:36:38 **dmit239.18-1 5 Pfam 6 сигнал** 140 База данных Pfam содержит информацию о белковых доменах. Домен имеет ID и AC и название, может указываться структура и функция(если известны). Профиль домена был составлен по seed alignment, ║ профиль и выравнивание в виде дерева также представлены в Pfam. Отдельно записаны доменные архитектуры с этим доменом и где они встречаются, распределение встречаемости по таксонам. Гомологичные домены из разных белков объ5диняются в семейства. Также в Pfam имеется информация о более коротких и менее эволюционно значимых единицах: мотивах, повторах и участках малой сложности в составе архитектуры. Некоторые белки, взаимодействующие с ДНК, например, рестриктазы или метилтрансферазы, требуют определённых последовательностей. Против таких последовательностей идёт отбор там, где активность белков не нужна. Для определения узнаваемого белком сигнала лучше всего подходит метод Chip-seq, когда мы можем выделить белок вместе со связанным с ним участком нуклеиновой кислоты, однако последовательностей и белком слишком много, чтобы с ними всеми работать [это же другая тема. Желание включить в текст как можно больше информации, по любому вопросу, порочно в науке]

**5/5** 10/04/2020 09:45:56 **daniil.bobrovsky 2 IC 120**

В теории информации для случайной величины определяется её информационное содержание I(x)=-logP(x): чем меньше вероятность события [события x], тем больше информации оно содержит. Матожидание информационного содержания называется информационной энтропией, H(x)=-P(x)logP(x).

Если мы построили выравнивание, то можем подсчитать частоту каждой буквы на каждой позиции. Тогда информационную энтропию символа i в позиции j можно найти как -p(i,j)\*log(p(i,j)), а информационное содержание (IC) выравнивания определить как сумму по всем символам и позициям. Теперь учтем, что частоты букв в геноме не равны, то есть наименее информативна для нас позиция, вероятности встретить все символы в которой равны вероятностям встретить их в среднем по всему геному. Тогда IC=-sum(p(i,j)\*log(p(i,j)/p(j))), где p(j) – частота буквы j в геноме. Чем выше информационное содержание выравнивания, тем меньше вероятность, что совпадение последовательностей в этом участке случайно."

**5/4** 510/04/2020 09:45:58 **kurkino 5 Pfam** 105

 Домен-единица белковой последовательности, в которой происходят локальные изменения, то есть для них имеет смысл делать филогенетическое дерево. База данных Pham - для семейств белковых доменов. Поиск по бд осуществляется за счет ID, AC или идентификатору домена.С помощью pham можно узнать доменную архитектуру белка, получить описание домена, 3d структуру домена и др. Семейства в бд формируются с помощью seed alignment, на этой основе строится профиль HMM. Далее происходит поиск по по бд \_\_\_ с определенным порогом. Полученная HMM есть в Pham.В pham можно узнать про кланы- семейства белковых доменов с общим эволюционным происхождением. С помощью картинки строения доменной архитектуры белка можно изучить:домены, повторы,трансмембранные участки и др.Некоторые домены представлены только частично(фрагментом).

**5/4** 10/04/2020 10:18:22 **osyafinkelberg 5 Pfam** 121 Доменом называется аминокислотная последовательность, которую можно найти во множественном выравнивании белков. Гомологичные домены из разных белков, - семейства доменов, - хранятся в базе данных Pfam.[две мысли: называются; хранятся]

Существует механизм пополнения Pfam новыми доменами (и семействами доменов), имеющимися в белках из UniProt. Он основан на статистической модели HMM (Hidden Markov Model):

Для имеющегося или вновь построенного множественного выравнивания (seed) специальная программа строит случайную последовательность, “похожую” на выравнивание. Эта программа задает HMM-профиль для данного семейства. Для каждой новой последовательности, найденной в UniProt, считается вероятность быть построенной данной программой. Если вероятность превосходит фиксированную пороговую величину, домен из последовательности включается в семейство. Пороговая величина определяется за счет калибровки HMM-профиля на случайном банке белковых последовательностей.

Бд Pfam делится на разделы Pfam-A и Pfam-B, которые курируются вручную и автоматически, соответственно.

===========================================================XXX Май >=23

**4/3** 24/05/2020; 00:**17**:16 **Скрипка 110**

PWM - инструмент для нахождения мотивов в выравнивании. Сперва подсчитываются частоты букв на каждой позиции выравнивания. Буквы могут в общем по геному отличаться по частоте, нужно отделить такие различия в частоте от вызванных наличием мотива. Наблюдаемая частота [Чего?] сравнивается с ожидаемой.

[Что на входе?] Технически PWM считается так: подсчитывается число каждой буквы на каждой позиции, к этим числам прибавляется псевдокаунт.[псевдокаунт это …] Полученные числа делятся на число последовательностей [в выравнивании, к которому прибавляется …], в которых есть позиция. Такая частота делится на ожидаемую, берётся логарифм. Псевдокаунты нужны чтобы не получать -inf при взятии логарифма от нуля если буква не встретилась. [не только]

Получается матрица значений не более 2 [нет] для каждой буквы в каждой позиции. Положительные значения значат повышенную встречаемость [Чего? Где?] , отрицательные – сниженную.

**10/05/2020 a.fokina 105 IC**

Информационное содержание выравнивания - сумма информационных содержаний всех его позиций. IC выравнивания визуализируется, например, с помощью LOGO, где высота букв соответствует информационному содержанию колонки.

Информационное содержание колонки - математическое ожидание веса в колонке, если распределение вероятностей встречи каждого нуклеотида в колонке задано частотами букв в ней. Чем больше значение информационного содержания выравнивания, тем сильнее различия между реально наблюдаемыми и ожидаемыми частотами букв, и тем больше информации содержится в колонке.

Вычисляется информационное содержание колонки следующим образом: сначала считается число различных букв в каждой колонке, потом частоты букв и частоты [???]умножаются на веса из позиционной весовой матрицы без псевдоотсчетов. Максимальное информационное содержание колонки равно весу из матрицы.

**5/4 28/05/2020; 00:04:56 a.fokina 100 IC**

Иногда бывает важно оценить количество информации, содержащееся в участке генома. Информационное содержание показывает, насколько необычна встреча конкретной буквы в конкретной позиции. Если информационное содержание буквы в этой позиции выше ожидаемого, можно сделать вывод, что она расположена там не случайно и участок выравнивания содержит важную информацию. Информационное содержание последовательности складывается из информационного содержания всех входящих в нее колонок. Для важных сигналов, закодированных в виде консервативных участков последовательности, информационное содержание будет большим. Информационное содержание вычисляется как логарифм отношения наблюдаемой частоты встречи буквы на данной позиции выравнивания к её средней частоте встречи в геноме с весом, равным наблюдаемой частоте буквы на конкретной позиции.

**5/4** 23/05/2020; 00:**27**:05 ivan.chernykh 96 сигналы и мотивы

Все биологические последовательности подверженны мутациям - через них осуществляется эволюция. Однако, при подробном рассмотрении становится ясно, что мутации не закрепляются на всех участках одинаково. Существуют специализированные участки, почти не подверженные изменчивости в силу того, что они играют важную биологическую роль, которая утратится даже при возникновении небольшого числа замен. Такие последовательности называются мотивами. Иногда мотивом называют целое объединение каких-то таких последовательностей. Есть отдельный тип мотивов, играющих очень важную роль. Этот тип ответственнен за "указание" сторонним молекулам каких-либо данных о другой молекуле. Например, промоторные участки обозначают место посадки рибосомы, тем самым образовывая старт трансляции. Такие мотивы называются сигналами.

**4/4** 30/05/2020 13:20:45; artemisbakulin PWM 73

ПВМ – это способ представления мотивов в биологических последовательностях. ПВМ содержит оценки встречаемости букв на точных [что значит точных?] позициях внутри мотива. Для составления ПВМ понадобится выравнивание последовательностей друг под другом. Для каждой колонки считают частоту букв (A, T, G, C для ДНК). Для того чтобы сделать поправку на встречаемость буквы в геноме вычисляют отношение правдоподобия: логарифм частного от деления наблюдаемой частоты буквы на ожидаемую. Консенсусную последовательность можно представить в виде logo, где вес отражается величиной буквы. [Нет, ЛОГО строится по инфрмационному содержанию]