Текст лекции Орлова Ю.Л. и Потапова В.Н. 26.04.2003
по марковским моделям с разбиением по слайдам.

(Слайд 1)
Уважаемые слушатели,
на данной лекции речь пойдет о
марковских моделях и их применении в биоинформатике.

(Слайд 2)
Все упомянутые формулы даны на слайдах как рисунки.

Краткое содержание лекции:

Теория марковских моделей.
Скрытые марковские модели.
Примения марковских моделей для предсказания функциональных районов в ДНК и белках.
Интернет-доступное программное обеспечение.

Различные разработки метода скрытых марковских цепей в настоящее время широко применяются для выравнивания последовательностей (белков и ДНК), а также для выявления гомологии между ними и поиска и распознавания последовательностей, некоторым обобщённым сходством. Более подробное изложение метода и способы его применения для анализа генетических текстов можно найти в книге (Durbin et al., 1998).
Под марковской цепью в обобщённом смысле подразумевается последовательность событий, каждое из которых происходит с определённой вероятностью.
Первые применения скрытых марковских моделей были для задачи распознавания речи (голоса человека)- Rabiner – speech recognition (1993)

Рекомендуемая литература по данной теме - учебник на английском языке по скрытым марковским моделям:
Durbin R., Eddy S.R., Krogh A., Mitchson G. Biological sequence analysis. 1998, Cambridge: Cambridge University Press, 356 p

(Слайд 3)
Основная задача биоинформатики, которую можно исследовать, используя математический аппарат марковских моделей, заключается в следующем. Предположим, у нас имеется два длинных генетических текста, нужно выяснить насколько эти тексты являются близкими по своей статистической структуре. Причём наш вывод должен быть устойчивым относительно малосущественных изменений текста, таких как вставки и выпадения отдельных символов, перестановка небольших фрагментов и т. д. Нужно сравнивать и небольшие генетические последовательности, но тогда их должно быть достаточно много. Статистическая близость последовательностей может оказаться следствием подобия их функциональных свойств или общности происхождения. Например, имея набор функциональных сайтов какого-либо типа можно выяснить, принадлежит ли к этому типу некоторая генетическая последовательность или нет.

(Слайд 4)
Вид марковской модели выбирается в зависимости от содержательной задачи. Ниже будут описаны два вида марковских моделей: контекстные и скрытые, а также будут указаны области их применения. Основная идея метода основывается на следующих соображениях. Пусть определёна марковская модель и некоторый набор исходных (обучающих) генетических текстов. Тогда по исходным данным определяются параметры

θ модели (конкретные способы будут описаны ниже). Вероятность Pr(X | θ) вновь предъявленной последовательности X при заданных параметрах θ определяет точность соответствия последовательности X параметрам модели. Чем больше величина Pr(X | θ), тем лучше последовательность X соответствует параметрам модели, тем больше оснований отнести последовательность X к тому же классу, что и исходные генетические тексты. Конечно, в действительности однозначно задана генетическая последовательность, а параметры модели мы выбираем по собственному произволу, поэтому нам важно насколько велика величина Pr(θ|X) – вероятность того, что параметры модели порождающей X равняются θ. Однако формула Байеса

показывает, что эти величины Pr(X | θ) и Pr(θ|X) прямо пропорциональны.

(Слайд 5)
Поскольку величины Pr(X | θ) обычно очень малы и существенно зависят от длины |X| последовательности X, то на практике обычно пользуются величиной зависящей от логарифма, которую можно интерпретировать, как сложность последовательности X в Марковской модели с параметрами θ, в расчёте на букву. Чем меньше сложность, тем лучше последовательность соответствует параметрам модели.

(Слайд 6)
Определение марковской модели

Пусть D – конечный **алфавит** и S – множество **состояний** модели. Рассмотрим ориентированный граф, вершинами которого являются состояния S, а рёбра помечены помечены буквами алфавита D.

Ориентированные пути по графу порождают слова в алфавите D. Говорят, что такой граф задаёт **конечный автомат,** а множество слов порождённых этим автоматом называют **регулярным языком**. Иногда среди состояний выбираются конечное и начальное состояния, т. е. состояния с которых начинается и заканчивается любое порождаемое слово.

(Слайд 7)

Одним из наиболее важных примеров графов определяющих марковскую модель является **граф де Брёйна**. Пусть , состояния и соединены ребром , если . Ребро соединяющее состояния σ и σ’ отметим буквой . На рисунке изображён граф де Брёйна при D={A,G,T,C}и S=D**2.

Приведен пример графа для последовательности ДНК:
atatctt

Заметим, что программа построения графа де Брёйна доступна в Интернете на сайте института математики СО РАН.

(Слайд 8)

Описанный способ порождения символьных последовательностей называется **марковской моделью,** если ввести вероятности P(σ|σ’,a) переходов между состояниями модели и вероятности P(a|σ) порождения букв в различных состояниях. Набор вероятностей называется набором параметров марковской модели. Параметры Модели должны удовлетворять следующим свойствам:

Вероятность пары состоящей из символьной последовательности и соответствующей ему последовательности состояний определяется рекуррентно из равенства (1)

(Слайд 9)
Как правило, рассматриваются два типа марковских моделей: контекстные и скрытые.

Контекстные марковские модели.

В контекстных марковских моделях предполагается, что наборы букв, которыми помечены стрелки, выходящие из одного состояния, не пересекаются. Так бывает когда в качестве состояний выбраны левые контексты букв в символьной последовательности. Тогда если нам известно предыдущее состояние и текущая буква, то однозначно определено и следующее состояние. Таким образом , если мы знаем начальное состояние и слово порождённое марковской моделью, то мы можем однозначно восстановить последовательность состояний, соответствующих этому слову. При этом возникают дополнительные условия на параметры марковской модели
При этом возникают дополнительные условия на параметры марковской модели
$P(\sigma|\sigma', a)= 1$, если ребро $(\sigma|\sigma')$ помечено буквой a и
$P(\sigma|\sigma', a)= 0$ в противном случае

Формула (1) приобретает вид

(Слайд 10)
Кроме того, величины $P(\sigma|\sigma')$ определяются равенствами
где сумма берётся по всем буквам, которыми помечена стрелка между состояниями σ и σ'. Из определений нетрудно заключить, что вероятность перейти в состояние σ зависит только от одного предыдущего состояния σ' и не зависит от других предшествующих состояний, т.е. последовательность состояний марковской модели представляет собой марковскую цепь.

Наиболее известным примером контекстных марковских моделей являются марковские модели конечного порядка t. В этом случае состояниями служат слова длины t, графом модели является соответствующий граф де Брёйна, а параметрами модели являются вероятности появления букв в различных контекстах.

Контекстную марковскую модель разумно использовать, когда исходные данные представляют собой небольшое число достаточно длинных текстов. Здесь мы вынуждены предполагать, что случайные последовательности, порождённые моделью с фиксированными параметрами **стационарны,** т.е. вероятность любой буквы зависит только от контекста, а не от номера позиции буквы в последовательности.
Свойство Марковской цепи (см.формулу).

(Слайд 11)
Выбор множества состояний модели, т.е. множества значимых с нашей точки зрения контекстов, вообще говоря, произволен. В то время как параметры марковской модели – вероятности букв в различных контекстах вычисляются из исходных данных по формулам

где r(w) – число включений слова w в исходные данные.
Например, пусть задана генетическая последовательность
GTAGTCTGATGCAT
На этом простом примере показано как
можно подсчитать условные вероятности букв A,T,G,C.

(Слайд 12)
Пример.

Реальной последовательности присваивается объединённая вероятность, вычисляемая по вероятностям отдельных событий. В марковской модели нулевого порядка, каждый следующий символ в последовательности не зависит от предыдущего (зависит от нулевого числа символов). Модель первого порядка определяет зависимость каждого символа только от одного предыдущего ему. Пример набора параметров марковской модели первого порядка приведён ниже.

Приведены вероятности символов и условные вероятности:
P(A) = 0.1, P(C)=0.3, P(G) = 0.2, P(T) = 0.4,
P(A|A) = 0.1, P(C|A)=0.3, P(G|A) = 0.2, P(T|A) = 0.4,
P(A|C) = 0.2, P(C|C)=0.1, P(G|C) = 0.4, P(T|C) = 0.3,
P(A|G) = 0.1, P(C|G)=0.2, P(G|G) = 0.3, P(T|G) = 0.4,
P(A|T) = 0.3, P(C|T)=0.1, P(G|T) = 0.4, P(T|T) = 0.2.

(Слайд 13)
Скрытые марковские модели.

Скрытые марковские модели естественно применять, когда имеется много генетических текстов небольшой длины. При этом предполагается, что все тексты начинаются с некоторого фиксированного состояния и вероятность буквы зависит в основном от номера позиции буквы в тексте.

Выделяется три вида состояний: основные, соответствующие позициям букв (их число примерно равно средней длине исходных генетических текстов); состояния-вставки и состояния-выпадения. Кроме того, одно основное состояние выделяется как начальное, а другое - как конечное. Основные состояния обозначаются квадратами, состояния-вставки обозначаются ромбами и состояния-выпадения обозначаются кругами.

(Слайд 14)

Этот тип марковских моделей подразумевает независимость порождения буквы в текущем состоянии и следующего состояния, т.е.

Таким образом, формула (1) приобретает вид (1')

Кроме того, в состояниях-выпадениях вообще не порождаются никакие буквы.
Как и в случае контекстных моделей, последовательность состояний модели является марковской цепью.

Эту марковскую модель называют скрытой, поскольку по последовательности букв нельзя однозначно восстановить последовательность состояний, в которых эти буквы порождались. Более того, каждой символьной последовательности может соответствовать множество путей на графе марковской модели.

(Слайд 15)
Для определения параметров скрытых марковских моделей нужно уметь решать следующую задачу. Пусть задана символьная последовательность и определены параметры модели, нужно найти последовательность состояний такую, что (см.формулу) вероятность этой последовательности состояний максимальна.
Или в логарифмической шкале - минус логарифм вероятности максимален.

Пользуясь последней рекуррентной формулой получаем искомые последовательности состояний.

Описанный метод носит название алгоритма Виттерби.

(Слайд 16)
На рисунке приведено формальное изложение алгоритма Витерби по (Durbin et al., 1998) Пример применения алгоритма для решения задачи о бросании костей (там же). Показана последовательность испытаний, соответствующая ей последовательность состояний и восстановление (предсказание) последовательности состояний по алгоритму Витерби.

(Слайд 17)
Выбор параметров скрытой марковской модели по исходным данным представляет собой некоторый итерационный процесс. Сначала определим параметры модели некоторым произвольным образом. Для каждой последовательности исходных данных с помощью алгоритма Витерби определи последовательность состояний , удовлетворяющую равенству (2). Теперь пересчитаем параметры модели по формулам

Затем для каждой последовательности исходных данных опять вычислим последовательности состояний и по указанным выше формулам определим новые параметры модели. Так повторяем до тех пор, пока процесс не стабилизируется, т.е. вновь вычисленные параметры не станут совпадать с прежними. Полученные в результате параметры скрытой марковской модели, вообще говоря, зависят от параметров выбранных первоначально.

(Слайд 18)
На слайде приведена скрытая марковская модель с определенными параметрами. Параметры приведены для множественного выравнивания нескольких сайтов, показанного в левой части слайда.
ACA---ATG
TCAACTATC
ACAC--AGC
AGA---ATC
ACCG--ATG

(Слайд 19)
Рассмотрим пример скрытой марковской модели другого типа
Использование марковских моделей для распознавания Вычисление переходных вероятностей a*st* для состояний метилирования CpG
Анализируется критерий отношения вероятностей данных в различных моделях: *model*+ и *model-*

(Слайд 20)
Смещение параметров.

При использовании указанных выше формул для вычисления параметров модели может оказаться, что некоторые последовательности вообще не могут порождаться моделью с такими параметрами из-за того, что вероятности отдельных букв и переходов между состояниями оказались равными нулю. Однако, равенство вероятностей нулю является статистически необоснованным, так как исходные данные всегда имеют ограниченный объём. Для недопущения нулевых вероятностей применяется **смещение** параметров.

(Слайд 21)
Марковская модель со смещенными параметрами.

(Слайд 22)

Марковские модели применяются как для ДНК, так и для белков.
Возможная скрытая марковская модель для аминокислотной последовательности ACCY. Белок представлен как последовательность вероятностей. Числа в квадратах показывают вероятность того, что аминокислота находится в данном состоянии, числа вдоль стрелок показывают переходные вероятности. Вероятность ACCY выделена жирными стрелками.

(Слайд 23)
Интернет-ресурсы по скрытым марковским моделям:
SAM HMM http://www.cse.ucsc.edu/research/compbio/sam.html
HMMER 2.2 http://hmmer.wustl.edu/
Марковская модель с переменной памятью
Complexity http://wwwmgs.bionet.nsc.ru/mgs/programs/complexity/

(Слайд 24)
На следующих слайдах приведены интерфейсы сайтов в Интернете по программам HMM и HMMER, являющимся хорошо проработанными ресурсами по использованию скрытых марковских моделей в биоинформатике.
Здесь доступ к программе SAM (Sequence Alignment and Modeling System)
http://www.cse.ucsc.edu/research/compbio/sam.html
программа разработана в университете Калифорнии (UCSC - University of California, Santa Cruz)
На этом сайте в Интернет также доступен учебный курс по скрытым марковским моделям на английском языке.

(Слайд 25)
Приведен интерфейс сайта университета Вашингтона в Сэнт-Льюисе с доступом к программе HMMER (произносится "Хаммер" - производное слово от английской аббревиатуры Скрытые Марковские Модели).

(Слайд 26-27)
Небольшой дополнительный комментарий
Условные вероятности определяются по правилу Байеса:
(См.формулу)
Данные вероятности получены согласно правилу перемножения вероятностей независимых событий. На практике (в случае более протяжённых последовательностей) обычно используют не саму эту вероятность, а её логарифм, то есть объединённая вероятность определяется не произведением, а суммой: log(a*b) = log(a) + log(b). Это позволяет избежать манипулирования бесконечно малыми величинами.

Рекомендуемая литература по теме скрытых марковских моделей:

Durbin R., Eddy S.R., Krogh A., Mitchson G. Biological sequence analysis. 1998, Cambridge: Cambridge University Press, 356 p.
Anders Krogh. An introduction to hidden Markov models for biological sequences. pp.45-63 //In: Salzberg S.L., Searls D.B., Kasif S. (eds), Computational Methods in Molecular Biology 1998 Elsevier Science B.V.