Марковские модели

Владимир Николаевич Потапов

Юрий Львович Орлов

Институт математики им.С.Л.Соболева СО РАН
Институт цитологии и генетики СО РАН

Структура лекции

(2)Теория марковских моделей
(3)Скрытые марковские модели
(4)Примения марковских моделей для предсказания функциональных райнов в ДНК и белках
(5)Интернет-доступное програмное обеспечение

Метод марковских цепей

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 применния 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 - Rabiner – speech recognition (1993)

Рекомендуемая литература: Durbin R., Eddy S.R., Krogh A., Mitchson G. Biological sequence analysis. 1998, Cambridge: Cambridge University Press, 356 p.

Введение

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.

```
>S416;
gaagccaccgggaaccacccatttcctccccatgtttgtcaagccgtCCTCAGGCgttga
cgacaacccctcacctcaaaaaactttcatggcacgcat
>S4252;
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnggatccagCCGCCACGCCtgtg
acactcgtgtgctttccctggtgtgtgcttgtggcaggtg
>S4254;
ttgtggcaggtgggggcagagggtcctcaggccagagagccactCCCCCAGCgccagacc
accctcttcctcactcccccacctcacccccctcacaggtg
```

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_0 модели (конкретные способы будут описаны ниже). Вероятность $Pr(X \mid \theta_0)$ вновь предъявленной последовательности X при заданных параметрах θ_0 определяет точность соответствия последовательности X параметрам модели. Чем больше величина $Pr(X \mid \theta_0)$, 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 $Pr(\theta_0 \mid X)$ – вероятность того, что параметры модели порождающей X равняются θ_0. Однако формула Байеса

$$Pr(\theta_0 \mid X) = \frac{Pr(X \mid \theta_0) Pr(\theta_0)}{\sum_\theta Pr(X \mid \theta) Pr(\theta)}$$

показывает, что эти ве

Поскольку величины $Pr(X \mid \theta_0)$ обычно очень малы и существенно зависят от длины $|X|$ последовательности X, то на практике обычно пользуются величиной

$$\frac{1}{|X|} \log \frac{1}{Pr(X|\theta_0)}$$

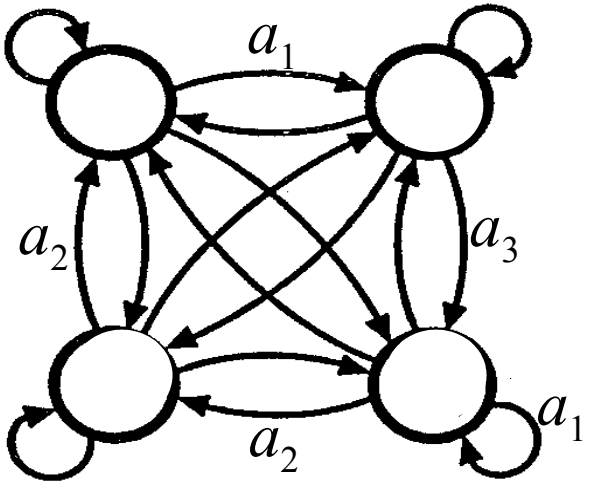
которую можно интерпретировать, как сложность последовательности X в марковской модели с параметрами θ_0, в расчёте на букву.

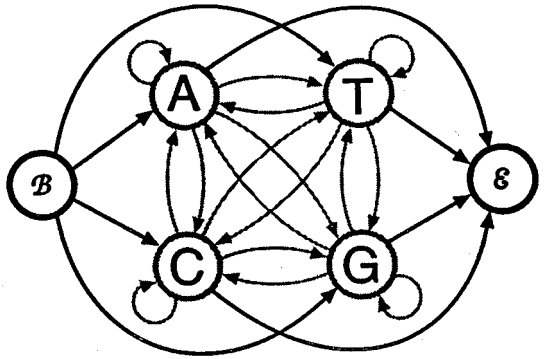
Чем меньше сложность, тем лучше последовательность соответствует параметрам модели.

Определение марковской модели

Пусть **D={a$_i$}** – 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.

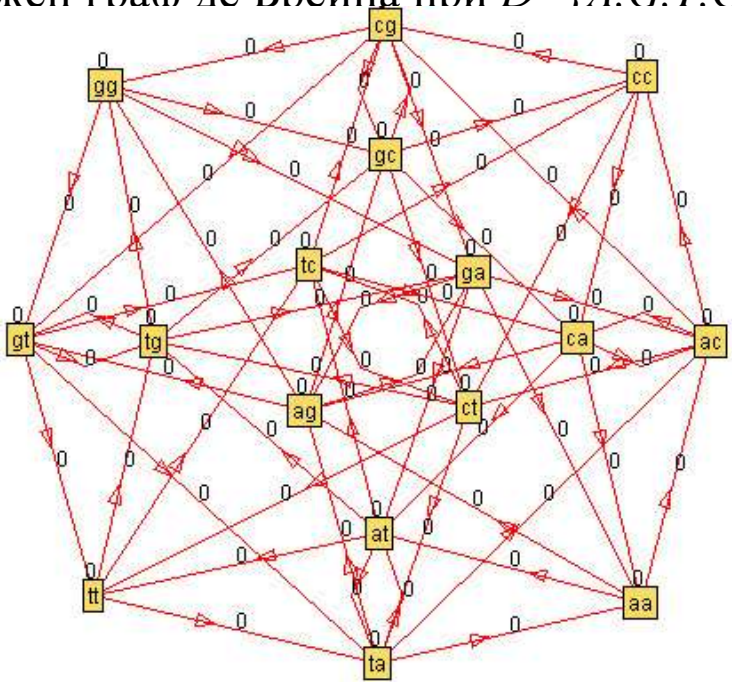
Говорят, что такой граф задаёт **конечный автомат,** а множество слов порождённых этим автоматом называют **регулярным языком**.

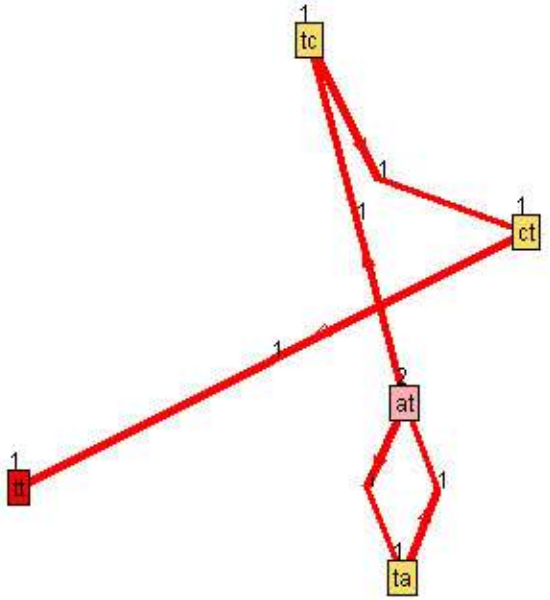
Иногда среди состояний выбираются конечное и начальное состояния, т. е. состояния с которых начинается и заканчивается любое порождаемое слово.

Одним из наиболее важных примеров графов определяющих марковскую модель является **граф де Брёйна**.

Пусть $S=D^t$, состояния $\sigma = a_{i_1} \ldots a_{i_t}$ и $\sigma' = a_{j_1} \ldots a_{j_t}$ соединены ребром, если $w = a_{i_2} \ldots a_{i_t}$ и $w = a_{j_1} \ldots a_{j_{t-1}}$ т.е. $\sigma = a_{i_1} w$ и $\sigma' = w a_{i_t}$.

Ребро соединяющее состояния σ и σ' отметим буквой a_{j_t}. На рисунке изображён граф де Брёйна при $D=\{A,G,T,C\}$ и $S=D^2$.





atatctt

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 следущим свойствам:

$$P(\sigma_j|\sigma_i, a_k) \geq 0, \sum_j P(\sigma_j|\sigma, a) = 1,$$

$$P(a_j|\sigma_i) \geq 0, \sum_j P(a_j|\sigma) = 1.$$

Вероятность пары состоящей из символьной последовательности $x^n=x_1x_2...x_n$ и соответствующей ему последовательности состояний $s^n=s_1s_2...s_n$ определяется рекуррентно из равенства

$$P(x^{n+1}, s^{n+1}) = P(x^n, s^n)P(s_{n+1}|s_n, x_n)P(x_{n+1}|s_{n+1}). \qquad (1)$$

Здесь $P(x^0, s^0) = P(\emptyset) = 1.$

Кроме того, нам потребуются формулы

$$P(x^n) = \sum_{s^n \in S^n} P(x^n, s^n), \qquad P(s^n|x^n) = \frac{P(x^n, s^n)}{P(x^n)}.$$

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

P(σ|σ', *a*)= 1, если ребро (σ|σ') помечено буквой *a* и

P(σ|σ', *a*)= 0 в противном случае.

Формула (1) преобразуется в

$$P(x^{n+1}) = P(x^{n+1}, s^{n+1}) = P(x^n, s^n)P(x_{n+1}|s_{n+1}).$$

Кроме того, величины P(σ|σ') определяются равенствами

$$P(\sigma|\sigma') = \sum_i P(a_i|\sigma')$$

где сумма берётся по всем буквам, которыми помечена стрелка между состояниями σ и σ'. Из определений нетрудно заключить, что вероятность перейти в состояние $\boldsymbol{s}_{i+1}$ зависит только от одного предыдущего состояния $\boldsymbol{s}_i$ и не зависит от других предшествующих состояний, т.е. последовательность состояний $\boldsymbol{s}^n$ 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.

$$P(x)=P(x_L, x_{L-1},\dots x_1) = P(x_L \mid x_{L-1},\dots x_1)\, P(x_{L-1} \mid x_{L-2},\dots x_1) \dots P(x_1)$$

Свойство Марковской цепи:

$$P(x)= P(x_L|x_{L-1})P(x_{L-1}|x_{L-2}) \dots P(x_2|x_1)\, P(x_1) = P(x_1) \prod_{i=2}^{L} a_{x_{i-1}x_i}$$

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

$$P(a_i|\sigma_j) = \frac{r(\sigma_j a)}{r(\sigma_j)},$$

где *r(w)* – число включений слова *w* в исходные данные.

Например, пусть задана генетическая последовательность

GTAGTCTGATGCAT

P(A|A)=P(AA)/P(A)=0/3=0.

P(T|A)=P(AT)/P(A)=2/3=0.66

P(G|A)=P(GA)/P(A)=1/3=0.33

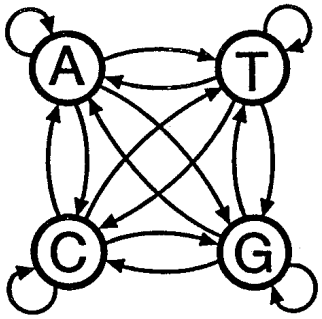
P(C|A)=P(GA)/P(A)=0/3=0.

Пример

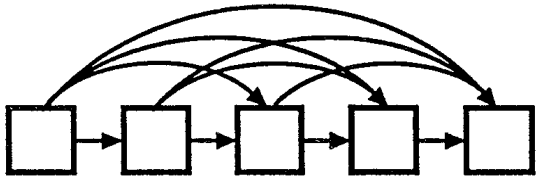
Реальной последовательности присваивается объединённая вероятность, вычисляемая по вероятностям отдельных событий. В марковской модели нулевого порядка, каждый следующий символ в последовательности не зависит от предыдущего (зависит от нулевого числа символов). Модель первого порядка определяет зависимость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

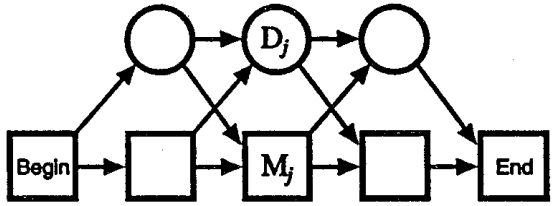
P(A) = 0.1,	P(C)=0.3,	P(G) = 0.2,	P(T) = 0.4,

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.



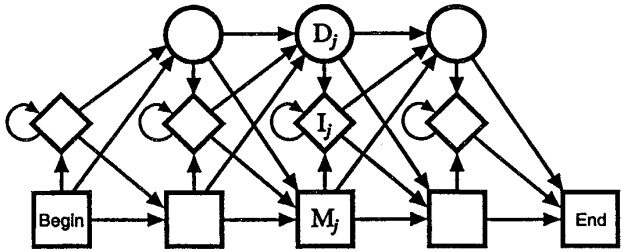
Скрытые марковские модели





Скрытые марковские модели естественно применять, когда имеется много генетических текстов небольшой длины (выровненные функциональные сайты). 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.



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

$$P(\sigma|\sigma', a) = P(\sigma|\sigma').$$

Таким образом формула (1) приобретает вид

$$P(x^{n+1}, s^{n+1}) = P(x^n, s^n)P(s_{n+1}|s_n)P(x_{n+1}|s_{n+1}). \qquad (1')$$

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.

Для определения параметров скрытых марковских моделей нужно уметь решать следующую задачу. Пусть задана символьная последовательность x^N и определены параметры модели, нужно найти последовательность состояний s^N такую, что

$$P(s^N(x^N)|x^N) = \max_{s^N} P(s^N|x^N) \qquad (2)$$

или по-другому $l(x^N, s^N(x^N)) = \min_{s^N} l(x^N, s^N)$,
где $l(\quad) = -\log P(\quad)$.

Из формулы (1') имеем равенство

$$l(x^{n+1}, s^{n+1}) = l(x^n, s^n) + l(s_{n+1}|s_n) + l(x_{n+1}|s_{n+1}), \qquad (3)$$

Для всех $\sigma \in S, n \geq 0$ определим $L(\sigma|x^{n+1}) = \min_{s^n} l(x^{n+1}, s^n\sigma)$.
Тогда из формулы (3) нетрудно найти

$$\min_{s^{n-1}\sigma'} l(x^{n+1}, s^{n-1}\sigma'\sigma) = \min_{\sigma'}\{\min_{s^{n-1}} l(x^n, s^{n-1}\sigma') + l(\sigma|\sigma') + l(x_{n+1}|\sigma)\},$$

$$L(\sigma|x^{n+1}) = \min_{\sigma'}\{L(\sigma'|x^n) + l(\sigma|\sigma') + l(x_{n+1}|\sigma)\}.$$

Пользуясь последней рекуррентной формулой получаем искомые

$$l(x^N, s^N(x^N)) \text{ и } s^N(x^N).$$

Описанный метод носит название алгоритма Виттерби

Формальное изложение алгоритма Витерби по (Durbin et al., 1998)

Пример применения алгоритма для решения задачи о бросании костей (там же)

Видны события (Rolls),

Реальные состояния

(Die – F,L) и

Решение по алгоритму Витерби

Algorithm: Viterbi

Initialisation ($i = 0$): $v_0(0) = 1,\ v_k(0) = 0 \text{ for } k > 0.$

Recursion ($i = 1 \ldots L$): $v_l(i) = e_l(x_i)\max_k(v_k(i-1)a_{kl});$
$\text{ptr}_i(l) = \text{argmax}_k(v_k(i-1)a_{kl}).$

Termination: $P(x, \pi^*) = \max_k(v_k(L)a_{k0});$
$\pi_L^* = \text{argmax}_k(v_k(L)a_{k0}).$

Traceback ($i = L \ldots 1$): $\pi_{i-1}^* = \text{ptr}_i(\pi_i^*).$

```
Rolls   315116246446644245311321631164152133625144543631656626566
Die     FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLLL
Viterbi FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLL

Rolls   651166453132651245636664631636663162326455236266666625151
Die     LLLLLLFFFFFFFFFFFFLLLLLLLLLLLLLLLLFFFLLLLLLLLLLLLLLFFFFFF
Viterbi LLLLLLFFFFFFFFFFFFLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLFFFFF

Rolls   222554416665665635643243641315134651463534111264146262 53
Die     FFFFFFFFLLLLLLLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
Viterbi FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF

Rolls   366163666466232534413661661163252562462255265252266435353
Die     LLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
Viterbi LLLLLLLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF

Rolls   233121625364414432335163243633665562466662632666612355245
Die     FFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLLLLLLLLLLLLFFFFFFFF
Viterbi FFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLLLLLLLLLLLFFFFFFFF
```

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 Виттерби определи последовательность состояний , удовлетворяющую равенству (2). Теперь пересчитаем параметры модели по формулам

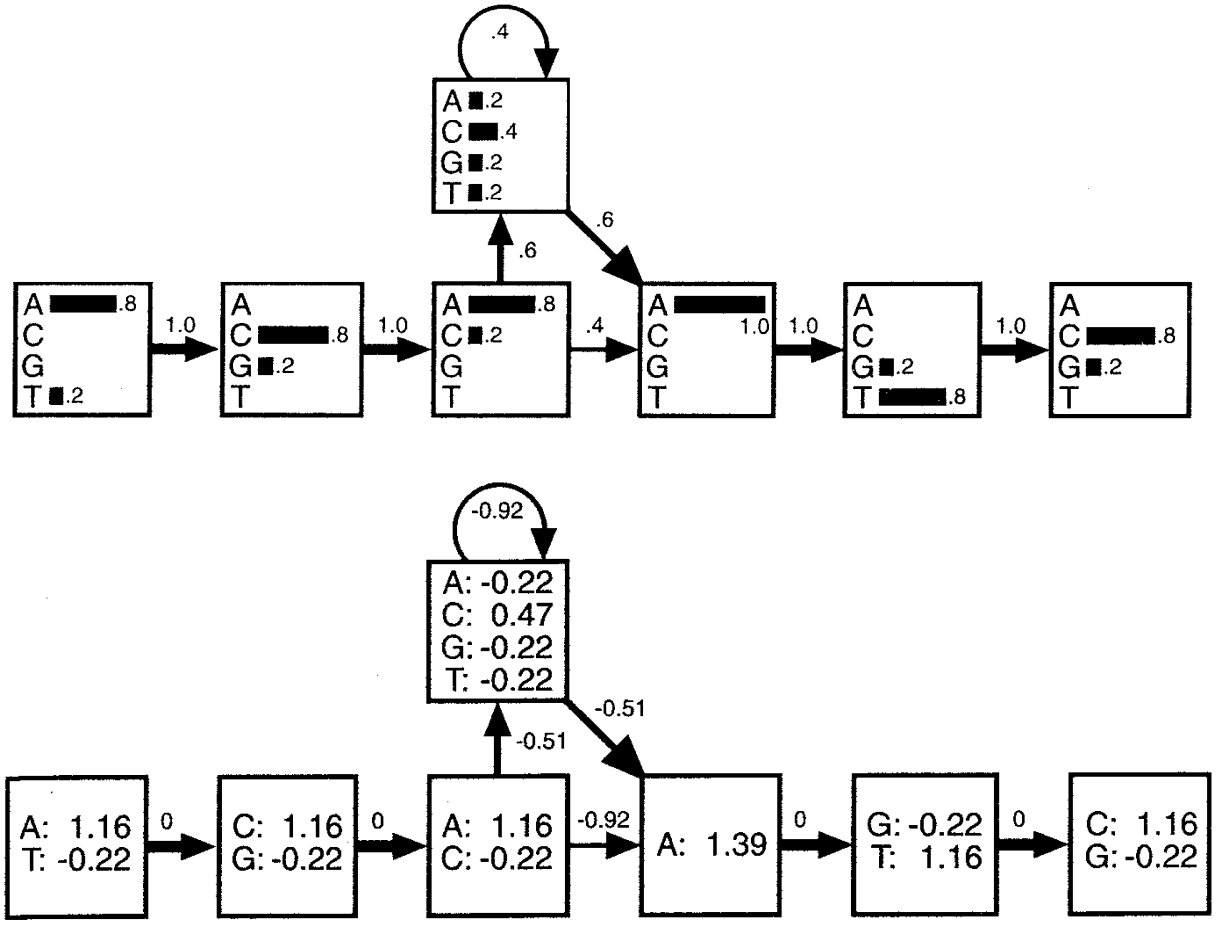
$$P(\sigma_i|\sigma_j) = \frac{r((\sigma_j, \sigma_i))}{r(\sigma_j)},$$

$$P(a_i|\sigma_j) = \frac{r(a_i, \sigma_j)}{r(\sigma_j)},$$

где $r((\sigma_j, \sigma_i))$ — число переходов из состояния σ_j в состояние σ_i в последовательности $s(X)$, $r(\sigma_j)$ — число состояний σ_j в последовательности $s(X)$, $r(a_i, \sigma_j)$ — число букв a_i, порождённых в состоянии σ_j в последовательности X.

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.

ACA---ATG
TCAACTATC
ACAC--AGC
AGA---ATC
ACCG--ATG

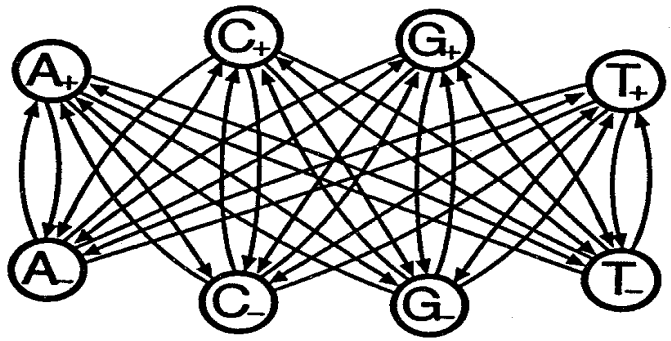


Скрытая марковская модель с определенными параметрами.

Пример скрытой марковской модели другого типа

Использование марковских моделей для распознавания

Вычисление переходных вероятностей a_{st} для состояний метилирования CpG



+	A	C	G	T
A	.180	.274	.426	.120
C	.171	.368	.274	.188
G	.161	.339	.375	.125
T	.079	.355	.384	.182

-	A	C	G	T
A	.300	.205	.285	.210
C	.322	.298	.078	.302
G	.248	.246	.298	.208
T	.177	.239	.292	.292

Анализируется критерий отношения вероятностей данных в различных моделях

$$S(x) = \log P(x|model+) / P(x|model-) =$$

$$= \sum_{i=1}^{L} \log a^{+}_{xi-1\,xi} / a^{-}_{xi-1xi} = \sum_{i=1}^{L} \log \beta_{xi-1\,xi}$$

Смещение параметров.

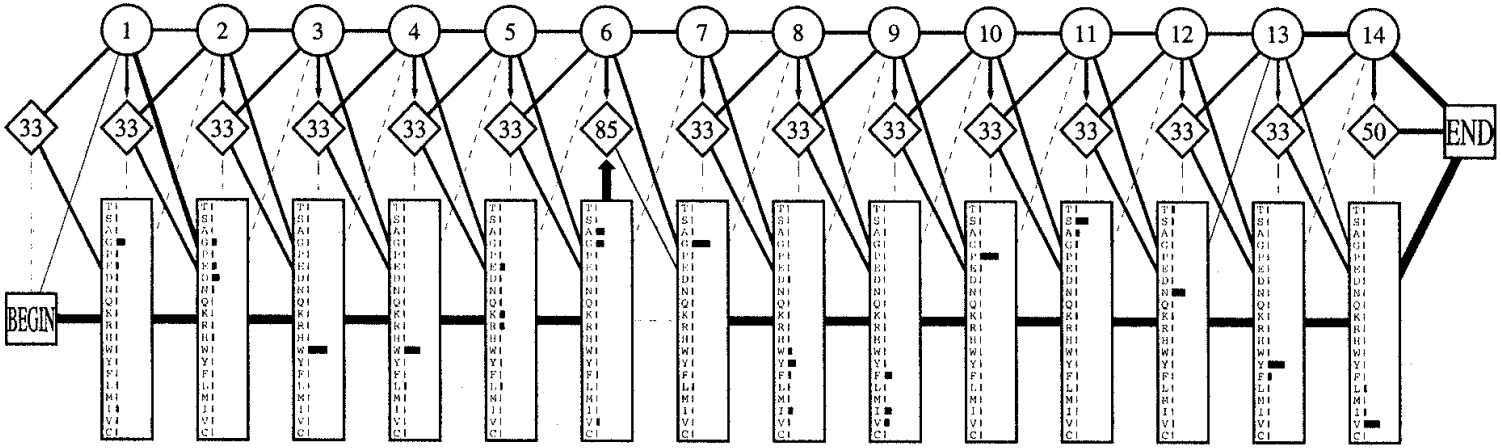
При использовании указанных выше формул для вычисления параметров модели может оказаться, что некоторые последовательности вообще не могут порождаться моделью с такими параметрами из-за того, что вероятности отдельных букв и переходов между состояниями оказались равными нулю. Однако, равенство вероятностей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

$$P(a_i|\sigma) = \frac{r_i + 1/2}{n/2 + \Sigma_i r_i}.$$

Очевидно $P(a_i|\sigma) > 0$ даже если $r_i = 0$ и равенство

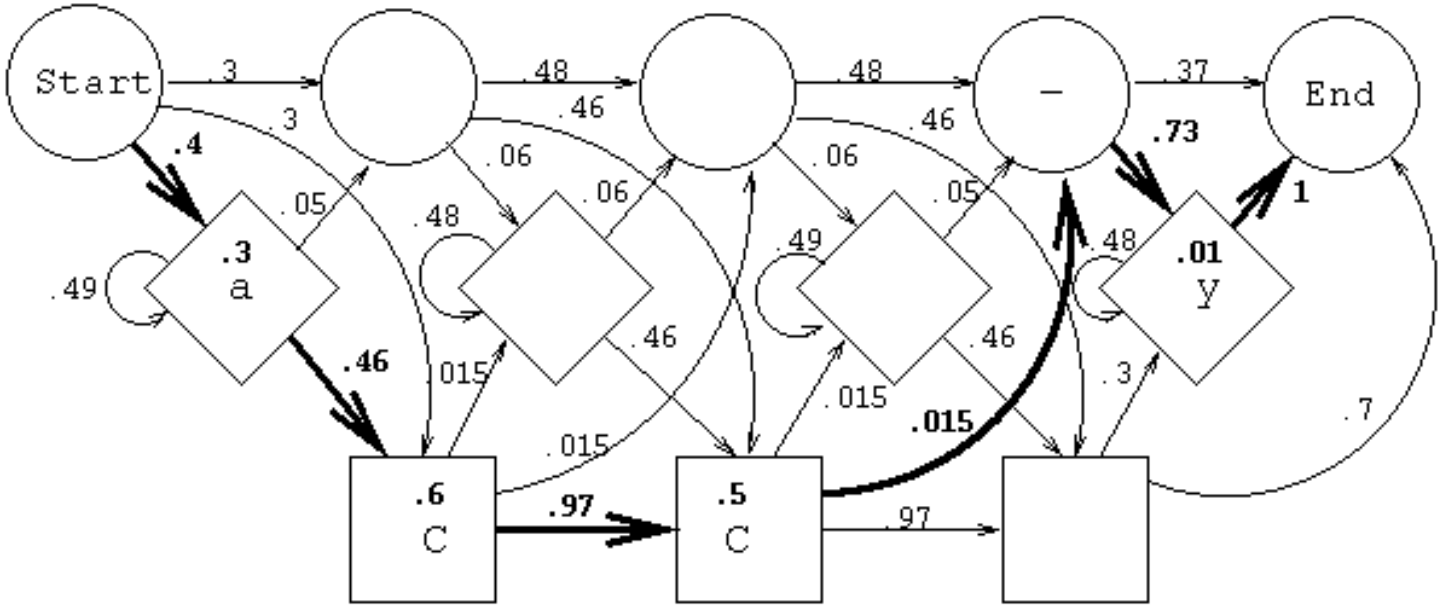
$$\sum_i P(a_i\sigma) = 1$$

сохраняется.



Марковская модель со смещенными параметрами.

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.

Интернет-ресурсы:

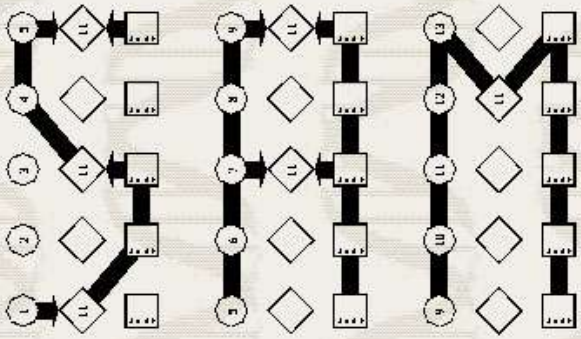
SAM HMM
http://www.cse.ucsc.edu/research/compbio/sam.html

HMMER 2.2 http://hmmer.wustl.edu/

Марковская модель с переменной памятью

Complexity
http://wwwmgs.bionet.nsc.ru/mgs/programs/complexity/

http://www.cse.ucsc.edu/research/compbio/sam.html



Sequence Alignment and Modeling System

SAM-T99 HMM WWW Servers

SAM 3.2 is available!
The SAM documentation (the 150+ page, 1.5+ MB manual is also available in PDF and PS) discusses the changes from previous versions.

If you are a college, university, U.S. government lab, or nonprofit, there is a free license to fill out and return by fax or hardcopy to UCSC. Otherwise, please request more information from sam-info@cse.ucsc.edu

Once you have the download information (or if you are only interested in our two graphical tools, hmmedit and sae), you can use the distribution page to obtain a copy of the software.

Martin Madera and Julian Gough have written a perl converter between SAM and HMMer 2.0 formats. You can get it from them (be sure to read their excellent documentation!) or download a 10/24/2000 copy.

Please read the ISMB99 tutorial on using HMMs

HMMER 2.2 http://hmmer.wustl.edu/

Washington University in St.Louis

Sean Eddy's Lab::HMMER
Department of Genetics
Washington University School of Medicine
St. Louis

| Dept. of Genetics | WashU | Medical School | Sequencing Center | CCB|
Eddy lab | Internal (lab only) | HMMER | PFAM | tRNAscan-SE | snoRNA database | Software | Publications |

HMMER 2.2

Profile hidden Markov models for biological sequence analysis

Profile hidden Markov models (profile HMMs) can be used to do sensitive database searching using statistical descriptions of a sequence family's consensus. HMMER is a freely distributable implementation of profile HMM software for protein sequence analysis.

A HMMER 2.2 beta release is now publicly available (5 August 2001). HMMER 2.2 is the first stable release of HMMER since December, 1998. 2.2 contains full support for the latest Pfam annotations; many new user-requested features; a number of small sensitivity and specificity enhancements; and a tiny, tiny number of bugfixes.

Documentation

- Text files associated with the HMMER 2.2g release:
  [README] [Installation] [Release notes] [License summary] [GNU General Public License]
- The HMMER User's Guide. [PDF, 92 pages]
- The theory behind profile HMMs: R. Durbin, S. Eddy, A. Krogh, and G. Mitchison,
  Biological sequence analysis: probabilistic models of proteins and nucleic acids,
  Cambridge University Press, 1998.
- Other publications from the Eddy group

Дополнительный комментарий

Условные вероятности определяются по правилу Байеса:

$P(G|T) = P(G,T)/ P(T)$,

здесь $P(G,T)$ – частота динуклеотида GT. Рассмотрим последовательность нуклеотидов GGTACC, для неё объединённая вероятности, построенные по моделям $0^{го}$ и $1^{го}$ порядка равны:

$P(G)* P(G)*..* P(C) = 0.2*0.2*0.1*0.4*0.3*0.3 = 1.44*10^{-4}$;

$P(G)* P(G|G)*..* P(C|C) = 0.2*0.3*0.4*0.3*0.3*0.1 = 2.16*10^{-4}$.

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.

$$S(X) = \log \frac{P(X \mid \text{модель}+)}{P(X \mid \text{модель}-)}$$

Hidden Markov models (HMMs)

Первые применения - Rabiner – speech recognition (1993)

Рассмотрим пример CpG island (Метилирование цитозина C в паре.)

$a_{st}=P(x_i=t|x_{i-1}=s)$ – вероятность перехода из одного состояния в другое

$$P(x)=P(x_L, x_{L-1},\dots x_1) = P(x_L \mid x_{L-1},\dots x_1)\, P(x_{L-1}|x_{L-2},\dots x_1) \dots P(x_1)$$

Свойство Марковской цепи:

$$P(x)= P(x_L|x_{L-1})P(x_{L-1}|x_{L-2}) \dots P(x_2|x_1)\, P(x_1) = P(x_1) \prod_{i=2}^{L} a_{xi\text{-}1xi}$$

Сумма вероятностей всех возможных последовательностей длины L:

$$\sum_{\{x\}} P(x)=\sum_{x1} \sum_{x2} \dots \sum_{x_I} P(x_1) \prod_{i=2}^{L} a_{xi\text{-}1xi}$$