

Дискуссии

УДК 519.217.2

И.В. СЕРГИЕНКО, А.М. ГУПАЛ

Институт кибернетики им. В.М. Глушкова НАН Украины,
пр. Академика Глушкова, 187, Киев, Украина
E-mail: d135@public.icyb.kiev.ua

СТАТИСТИЧЕСКИЙ АНАЛИЗ ГЕНОМА



А.А. Марков исследовал литературные тексты и показал, что вероятность гласной и согласной зависит от одной или двух предыдущих. Молекула ДНК записана в четырехбуквенном алфавите нуклеотидов. Вычисления на ЭВМ показали, что частоты букв и оценки переходных вероятностей стабильны на всех 24 хромосомах генома. Структура ДНК — однородная и эргодическая цепь Маркова, и для нее выполняется закон больших чисел. Регулярность цепи — основа процесса репликации ДНК, частоты букв должны быть стабильными по длине цепочки и соответствовать концентрации нуклеотидов водного раствора.

© И.В. СЕРГИЕНКО, А.М. ГУПАЛ, 2004

Введение. В 2003 г. было объявлено о завершении программы «Геном человека». Установлено, что последовательность ДНК человека состоит из трех миллиардов пар оснований и включает около 40 тысяч генов [1, 2]. Более половины размера генома занимают различные виды повторяющихся последовательностей, их роль пока непонятна, как не выяснены особенности функционирования генома в целом и каким образом отдельные его участки влияют на функции организма человека. Последовательности, кодирующие белки, составляют менее 1 % объема генома.

Генетические свойства организмов зачастую являются результатом многих недостаточно известных процессов и механизмов. Из-за сложности причинно-следственных связей их изучение с помощью существующих детерминированных математических моделей невозможно. В обзоре [3] отмечается, что многие генетические свойства особей, популяций или видов — следствие закономерностей, которые по своей сути являются стохастическими и их невозможно исследовать без использования *вероятностных моделей*. В последние годы возрос интерес к байесовским методам в биологии и генетике. Этот фактор можно объяснить тем, что байесовские индуктивные процедуры, построенные на основе наблюдаемых данных, являются оптимальными на таких структурах, как цепи Маркова, байесовские сети и независимые признаки. Они имеют полиномиальные оценки точности в зависимости от размеров наблюдаемых выборок и количества признаков изучаемых объектов [4, 5].

Для того чтобы проводить исследования с помощью известных статистических методов на таком большом массиве данных о структуре генома, нужно построить и исследовать его математическую модель. Один из возможных подходов к решению этой проблемы изложен в работе [6], где проведен детальный статистический анализ цепей Маркова. Для больших выборок исследовано поведение оценок переходных вероятностей, построенных в виде частот, и разработаны принципы определения модели цепи на основе проверок статистических гипотез.

Исходя из этих результатов и учитывая тот факт, что хромосомы человека имеют большую длину (порядка 10^8 нуклеотидов), можно предположить, что однородная цепь Маркова

лучше всего соответствует анализу установления закономерностей построения и функционирования генома. Частоты четырех букв и оценки переходных вероятностей стабильны для ДНК всех хромосом человека и других высших организмов.

Поскольку наша гипотеза опирается на результаты Маркова [7], коротко остановимся на его схеме рассуждений.

Анализ литературных текстов. В начале XX века Марков опубликовал ряд работ, которые привели к общей схеме «испытаний, связанных в цепь». На этой схеме ученый установил ряд замечательных закономерностей, положивших начало современной теории марковских процессов.

А.А. Марков рассмотрел поучительные примеры связанных испытаний и показал, что если рассматривать буквы литературного текста, то вероятность гласной и согласной зависит от одной или двух предыдущих букв. Наиболее известный результат А.А. Маркова — теорема эргодичности. Проведенный им статистический анализ литературных текстов подтвердил наличие в них свойств однородности и эргодичности для связанных цепей.

Рассматривается неограниченный ряд испытаний или измерений, которые отмечаются по порядку номерами

$$1, 2, \dots, k, k+1, \dots$$

Испытания связаны в *однородную цепь*: если для одного из испытаний появилось событие Γ (гласная буква) или противоположное ему событие C (согласная буква), то следующие за ним испытания зависят от этого результата, но не зависят от результатов предшествующих ему испытаний (в общей схеме испытания могут зависеть от нескольких предыдущих результатов).

Для всей цепи определены одни и те же два числа — p_1, p_2 . Число p_1 означает вероятность события Γ при $k+1$ -м испытании, если дано, что Γ появилось при k -м испытании. Число p_2 означает вероятность события Γ при $k+1$ -м испытании, если k -е испытание вызвало событие C .

Чтобы сообщить выводам полную определенность и уметь вычислять вероятности событий Γ и C при любом испытании, следует ввести еще число p^l , представляющее вероятность события Γ при первом испытании.

Заметим, что в двухбуквенном алфавите однородная цепь определяется тремя вероятностями, поскольку вероятности пар $\Gamma C, CC$ соответственно равны $1-p_1$ и $1-p_2$.

А.А. Марков первым изучил эргодические свойства цепей. Исследовался ряд чисел

$$p^l, p^2, \dots, p^k, p^{k+1}, \dots,$$

соответственно представляющих вероятности события Γ (быть букве гласной) при каждом из испытаний $l, 2, \dots, k, k+1, \dots$

Введя дополнительные параметры δ и p , определяемые равенствами $\delta = p_1 - p_2, p_2 = p(1 - \delta)$, и считая $|\delta| < 1$, он вывел общую формулу

$$p^k = p + (p^l - p) \delta^{k-l}, \quad (1)$$

откуда видно, что число

$$p = \frac{p_2}{1 + p_2 - p_1} \quad (2)$$

служит пределом $p^k \rightarrow p, k \rightarrow \infty$.

Заметим, что сходимость в формуле (1) — геометрическая и предел p не зависит от начальной вероятности p^l . Из теоремы эргодичности вытекает, что однородная цепь является слабо зависимой последовательностью, так как влияние значения начальной буквы на последующие быстро убывает при удалении от начала последовательности.

А.А. Марков рассмотрел последовательность 20 000 букв в романе А.С. Пушкина «Евгений Онегин», она составляет 20 000 зависимых испытаний, каждое из которых дает гласную или согласную букву.

Литературный текст поэмы А.С. Пушкина — осмысленная и зависимая последовательность 32 букв русского алфавита. Двухбуквенная последовательность гласных и согласных букв, выделенная из этого фрагмента литературного текста, никакого явного смысла не имеет.

А.А. Марков допустил существование неизвестной постоянной вероятности p — быть букве гласной и приближенную величину числа p он нашел из наблюдений, считая число появившихся гласных букв в 200 последовательностях по 100 букв. Полученное таким способом значение p оказалось равным 0,4319. Кроме числа p , А.А. Марков нашел также из наблюдений приближенные величины двух других чисел p_1 и p_2 , представляющих вероятности: $p_1 =$

гласной букве следовать за гласной, p_1 — гласной букве следовать за согласной.

При вычислении вероятностей p_1 и p_2 подсчитывается число пар (гласная, гласная), которое при делении на число всех гласных в тексте дает для p_1 приближенную величину $p_1 \approx 0,128$. Аналогично получилось, что $p_2 \approx 0,663$.

Отсюда вытекает, что вероятность букве быть гласной или согласной существенным образом зависит от того, какова предыдущая буква. Знаменатель прогрессии в формуле (1) равен $\delta = -0,535$ и $|\delta|^{11} \approx 0,001$.

Подставив полученные значения p_1 и p_2 в формулу (2), получим число 0,4319, совпадающее с ранее полученным приближенным значением для вероятности p . Нетрудно показать, что для однородной цепи (при вычислении переходных вероятностей p_1 и p_2 в виде частот) предел p в (2) совпадает с частотой гласных букв в цепочке текста. На основе проведенных расчетов мы видим, что эта особенность имеет место для литературного текста. Таким образом, вероятность букве быть гласной становится постоянной и равной p по всей длине текста (в силу быстрой сходимости в (1)). В этом смысле двухбуквенный текст является регулярной последовательностью.

Для эргодической цепи справедлив закон больших чисел [7], из которого вытекает, что частота гласной буквы длинной цепочки текста мало отличается от значения p в формуле (2). В нашей ситуации совпадение подсчитанной частоты гласной буквы с вероятностью p служит примером действия этого принципа. Подобное исследование А.А. Марков выполнил над произведением прозы другого автора.

Статистический анализ генома. Поведение однородной связанный цепи применительно к геному определяется начальным распределением вероятностей четырех букв: $p(A)$, $p(C)$, $p(G)$, $p(T)$, составляющих в сумме 1, и переходных вероятностей, записанных в виде матрицы

$$P = \begin{vmatrix} p(AA) & p(AC) & p(AG) & p(AT) \\ p(CA) & p(CC) & p(CG) & p(CT) \\ p(GA) & p(GC) & p(GG) & p(GT) \\ p(TA) & p(TC) & p(TG) & p(TT) \end{vmatrix},$$

где суммы вероятностей по строкам также равны единице. Переходные вероятности $p(ij)$, $i, j \in \{A, C, G, T\}$ имеют такой же смысл, как и у А.А. Маркова.

Одной из основных особенностей в анализе нуклеотидных последовательностей является то, что частоты встречаемости соседних оснований не являются независимыми. В частности, частоты пар соседних оснований обычно отличаются от произведений частот самих оснований. Иными словами, если p_u — частота основания типа u в последовательности и p_{uv} — частота, с которой соседние основания принадлежат к типам u и v , то

$$p_{uv} \neq p_u p_v.$$

Эта особенность частот была проанализирована в [8], где также изучался вопрос соответствия между марковской цепью определенного порядка и отдельными последовательностями ДНК. Отмечалось, что анализ марковских цепей следует проводить на уровне *всего генома*, а не на уровне отдельного гена.

Марковская цепь порядка k предполагает, что нахождение основания в определенном месте последовательности зависит только от оснований, находящихся в предшествующих k положениях. В цепи порядка 1 вероятность нахождения какого-либо основания в позиции i зависит от вероятности присутствия одного из четырех оснований в позиции $i-1$. Последовательность, состоящая из независимых оснований, соответствует марковской цепи 0-го порядка. Порядок цепи может быть установлен методами правдоподобия (на основе вероятностей нуклеотидной цепочки) путем решения серии задач распознавания гипотез.

В работе [6] показано, что для больших выборок величина

$$-2 \ln \frac{L(k)}{L(k+1)}$$

подчиняется известному распределению χ^2 , где $L(k)$ — правдоподобие цепи порядка k . Для цепи порядка k число независимых параметров составляет $3 \cdot 4^k$. Число степеней свободы статистики χ^2 равно разности между числами степеней свободы для каждого из двух порядков.

Проведенные расчеты на последовательностях генома показали, что существенное различие (на 5%-ном уровне) имеется только между цепями порядка 0 и 1.

При оценке порядка цепи k , при котором достигается наилучшее соответствие, следует иметь в виду, что цепи более высокого поряд-

ка имеют большое число степеней свободы. Поэтому целесообразно ввести «штраф» при слишком большом числе учитываемых параметров [8]. Для этого предлагается использовать информационный критерий Байеса (BIC):

$$BIC(k) = \text{Const} - 2\ln L(k) + 3 \cdot 4^k \ln n_k,$$

где n_k — число подпоследовательностей длины $k+1$, находящихся в рассматриваемой последовательности. То значение k , для которого $BIC(k)$ минимально, дает оценку порядка цепи. Проведенные расчеты на геноме человека подтверждают вывод о том, что цепь порядка 1 лучше всего соответствует данным.

Оценки переходных вероятностей $p(ij)$ вычисляются по формуле

$$\hat{p}(ij) = \frac{m(ij)}{\sum_j m(ij)}, \quad (3)$$

где $m(ij)$ — число пар (ij) . Заметим, что знаменатель по сути совпадает с $m(i)$ (число букв i в хромосоме).

Основная трудность исследования оценок $\hat{p}(ij)$ состоит в том, что знаменатель в формуле (3) не фиксирован, как в случае независимых испытаний, и оценки $\hat{p}(ij)$ являются смешенными. В работах [5,6] проведен статистический анализ оценок $\hat{p}(ij)$ при $m \rightarrow \infty$, где m — длина цепи. Величины $\hat{p}(ij) - p(ij)$ имеют предельное нормальное распределение со средним 0, дисперсиями

$$\frac{p(ij)(1-p(ij))}{mp(i)},$$

и для четырех различных значений $i \in \{A, C, G, T\}$ эти величины асимптотически независимы. Хромосомы имеют длину порядка 10^8 букв, поэтому дисперсии оценок малы.

Цепь Маркова оказывается простой и экономной моделью исследования генома. Для оп-

ределения вероятностей нуклеотидных последовательностей необходимо задание 15 вероятностей (3 — для начального состояния и 12 — для переходных вероятностей). Вычисления показали, что оценки этих вероятностей, построенные в виде частот, стабильны на всех 24 хромосомах генома человека (исследовались также оценки переходных вероятностей в зависимости от одной до четырех предыдущих букв). Частоты букв A и T составляют число 0,29, а для букв C и G — 0,21. Удивительно, что природа использовала одну и ту же схему записи информации для всех хромосом ДНК. Ведь вполне могло быть так, что отдельные хромосомы могли записываться с разными переходными вероятностями.

В табл. 1 приведены частоты букв и частоты пар букв (оценки переходных вероятностей) в хромосоме I генома человека.

Оценки переходных вероятностей для всех 24 хромосом строго положительны и находятся в диапазоне 0,05–0,34. Как и в двухбуквенном алфавите, это означает, что такая цепь является эргодической [9]: при $n \rightarrow \infty$ $\hat{p}(ij)^n$ сходится к предельным значениям $(\pi_A, \pi_C, \pi_G, \pi_T)$, не зависящим от i и образующим распределение вероятностей $\pi_A + \pi_C + \pi_G + \pi_T = 1$. Здесь $\hat{p}(ij)^n$, $i, j \in \{A, C, G, T\}$ — элементы матрицы \hat{P}^n , составленной из оценок переходных вероятностей, причем сходимость также происходит с геометрической скоростью. Имеет место интересная особенность поведения предельных значений π_j . Если переходные вероятности $p(ij)$ вычисляются в виде частот (3), то значения π_j , $j \in \{A, C, G, T\}$ совпадают со значениями частот букв j .

Вычисления на ЭВМ показали, что для всех хромосом имеет место сходимость величин $\hat{p}(ij)^n$ к предельному вектору $(\pi_A, \pi_C, \pi_G, \pi_T)$ примерно за 50 итераций, и этот вектор мало отличается

Частоты букв и пар букв в хромосоме I генома человека

Таблица 1

Частоты букв	Частоты пар букв			
A — 0,29137	AA — 0,32685	AC — 0,17234	AG — 0,24438	AT — 0,25643
C — 0,20838	CA — 0,34876	CC — 0,26056	CG — 0,04823	CT — 0,34245
G — 0,20834	GA — 0,28666	GC — 0,21137	GG — 0,26047	GT — 0,24150
T — 0,29190	TA — 0,21836	TC — 0,20497	TG — 0,24945	TT — 0,32722

от значений частот букв A, C, G, T . Вероятность отдельной буквы из множества $\{A, C, G, T\}$ достаточно быстро становится постоянной при удалении от начала цепочки. Таким образом, для текста генома, как и для двухбуквенной последовательности Маркова, выполняется свойство эргодичности. В табл. 2 приведены частоты чередования букв для трех геномов высших организмов, из которой видно, что у геномов других высших организмов такая же схема записи информации, что и в геноме человека. В таком случае не столь фантастической выглядит проблема спонтанной сборки белков.

Проблема сборки белков. Живая природа развивается эволюционным путем, т.е. индуктивно от простого к сложному. Изучение структур генов геномов человека и ряда других организмов показало в них наличие модификаций и различного рода перестановок генетического материала. Родственные гены обнаружены в геномах представителей всех трех царств живых организмов, причем замечено, что такие эволюционные процессы не прекращаются и протекают постоянно [10]. Перестройки и изменения генов проводятся в некодирующих участках генома. Белок-кодирующие гены, наоборот, весьма стабильны и имеют древнее происхождение. В ходе эволюции они меняются медленнее, чем окружающие их некодирующие участки генома. На наш взгляд, эта особенность поведения генома позволяет прояснить феномен самосборки белков.

Для записи белка, состоящего из 100 аминокислот, необходимо 300 нуклеотидов из алфавита A, C, G, T . Чтобы найти белок указанной длины, нужно перебрать $4^{300} = 2^{600}$ число различных вариантов, которое превышает 10^{180} . Такое число представить и сравнить с чем-либо невозможно (число атомов во Вселенной оценивается как 10^{70}). Удивительно, каким способом природа сумела создать за сравнительно короткий промежуток времени (порядка 1 млрд лет) такое большое количество функциональных белков.

Если учесть объем воды на Земле, то оказывается, что за такое время случайным перебором можно построить белок не более чем в 60–70 нуклеотидов, т.е. 25 аминокислот. Аналогичные расчеты случайного синтеза ДНК приведены в [11]. Учитывая отмеченную особенность

Таблица 2
Частоты чередования букв для трех геномов
высших организмов

Частоты пар букв	Человек		Мышь		Крыса	
	Хромосомы					
	1	2	1	2	1	2
AA	0,33	0,33	0,32	0,31	0,31	0,32
AC	0,17	0,17	0,18	0,18	0,19	0,18
AG	0,24	0,23	0,25	0,26	0,26	0,24
AT	0,26	0,26	0,26	0,25	0,25	0,26
CA	0,35	0,36	0,36	0,36	0,35	0,36
CC	0,26	0,25	0,25	0,25	0,26	0,24
CG	0,05	0,05	0,04	0,04	0,05	0,04
CT	0,34	0,35	0,35	0,35	0,35	0,35
GA	0,29	0,29	0,30	0,29	0,29	0,31
GC	0,21	0,21	0,19	0,20	0,20	0,19
GG	0,26	0,25	0,25	0,25	0,25	0,24
GT	0,24	0,25	0,26	0,25	0,25	0,26
TA	0,22	0,23	0,22	0,22	0,21	0,22
TC	0,20	0,20	0,21	0,21	0,22	0,21
TG	0,25	0,24	0,25	0,26	0,26	0,25
TT	0,33	0,33	0,32	0,31	0,31	0,32

поведения отдельных генов в различных геномах, можно сделать вывод о том, что белки высших организмов строятся из блоков коротких фрагментов, поскольку образовать длинные последовательности спонтанным перебором невозможно. В таком случае вся комбинация нуклеотидов также является марковской цепочкой.

Выводы. Согласно известной модели Уотсона-Крика молекула ДНК состоит из двух комплементарных цепочек. Поэтому нечетный трехбуквенный алфавит сразу же исключается из рассмотрения. В двухбуквенном алфавите размер генома увеличился бы в два раза, а это с учетом огромной длины ДНК является серьезным ограничением.

Статистический анализ и расчеты на ЭВМ показывают, что нуклеотидный состав хромосом соответствует модели однородной цепи Маркова и для нее выполняется эргодическое свойство, которое означает, что вероятности отдельных букв достаточно быстро становятся

Статистический анализ генома

постоянными при удалении от начала текста. Легко заметить, что такая модель экономна и наименее чувствительна к ошибкам в линейной записи. При замене всего лишь одной буквы текста ДНК в цепях более высокого порядка появляется большое число ошибок при определении переходных вероятностей. На наш взгляд, эргодичность цепи — основа процесса репликации ДНК, поскольку частоты нуклеотидов должны быть стабильны по длине цепочки ДНК и соответствовать концентрации нуклеотидов водного раствора.

SUMMARY. The chemical structure of DNA is characterized by sequences of four basic nitrogenous occurring in one of two nucleic acid chains and in a complementary fashion in the other. Markov chain is the aspect of probability theory that analyzes discrete states in which transition is a fixed probability not affected by the history of the system. It is shown that DNA is represented in the form of regular Markov chain. Ergodicity property and law of large numbers follow from the statistical analysis of stationary transition probabilities.

РЕЗЮМЕ. А.А. Марков дослідив літературні тексти і з'ясував, що ймовірність голосної та приголосної залежить від одної чи двох попередніх. Чотирьохлітерний текст ДНК записано у вигляді однорідного ланцюга Маркова, який є слабкозалежною послідовністю. Статистичний аналіз переходних ймовірностей виявив, що ланцюг є ергодичним та виконується закон великих чисел. Звідси випливає ряд важливих особливостей генома. Ергодичність ланцюга — основа про-

цесу реплікації ДНК, частоти нуклеотидів повинні бути стабільними уздовж послідовності ДНК та відповідати концентрації нуклеотидів водного розчину.

СПИСОК ЛИТЕРАТУРЫ

1. Lander E.S. et al. Human Genome Sequencing Consortium // Nature. — 2001. — **409**. — P. 860–921.
2. Venter J.C. et al. The sequence of the human genome // Science. — 2001. — **291**. — P. 1304–1351.
3. Beaumont M.A., Rannala B. The Bayesian revolution in genetics // Nature Rev. Genet. — 2004. — **5**. — P. 251–261.
4. Гупал А.М., Пашко С.В., Сергиенко И.В. Эффективность байесовской процедуры распознавания // Кибернетика и системный анализ. — 1995. — № 4. — С. 78–89.
5. Гупал А.М., Вагис А.А. Статистическое оценивание марковской процедуры распознавания // Проблемы управления и информатики. — 2001. — № 2. — С. 5–15.
6. Anderson T.W., Goodman L.A. Statistical inference about Markov Chains // Ann. Math. Statist. — 1957. — **28**. — P. 89–110.
7. Марков А.А. Исчисление вероятностей. — М., 1924. — 592 с.
8. Вейр Б. Анализ генетических данных. — М.: Мир, 1995. — 400 с.
9. Ширяев А.Н. Вероятность. — М.: Наука, 1989. — 640 с.
10. Сингер М., Берг П. Гены и геномы : В 2 т. — М.: Мир, 1998. — Т. 2. — 392 с.
11. Чернавский Д.С. Синергетика и информация. — М.: УРСС, 2004. — 288 с.

Поступила 16.03.04