УДК 577.21

БИОФИЗИЧЕСКИЕ ОСНОВЫ ОРГАНИЗАЦИИ ГЕНОМА

BIOPHYSICAL BASES OF GENOME ORGANIZATION

Волобуев А.Н.
Петров Е.С.
Романчук Н.П.

ФГБОУ ВО «Самарский государственный медицинский университет» Минздрава России

Volobuev AN
Petrov ES
Romanchuk NP

Samara State Medical University

Цель — анализ нуклеотидных последовательностей молекул ДНК, основ хранения информации с помощью ДНК.

Материалы и методы – использование теории цепей Маркова и информационного критерия Байеса.

Результаты. Исследованы принципы построения генетического кода. Показана целесообразность вырождения генетического кода. На основе применения теории цепей Маркова дан анализ конкретных нуклеотидных последовательностей. Рассмотрен метод секвенирования нуклеотидной последовательности.

Заключение. На последовательность нуклеотидов накладываются ограничения, связанные с комплементарностью оснований вдоль цепи ДНК. Эти ограничения на уровне последовательности кодонов могут в значительной мере сниматься вырожденностью генетического кода.

Ключевые слова: геном, нуклеотидная последовательность, цепи Маркова, информационный критерий Байеса, секвенирование.

Aim – the analysis of nucleotide sequences of DNA molecules and the bases of the information storage with the help of DNA.

Material and methods – the study is based on Markov chain theory and Bayesian Information Criterion.

Results. Principles of genetic code construction were investigated. Specific nucleotide sequences were analyzed using Markov chain theory; the method of sequencing nucleotide sequences was described.

Conclusion. A nucleotide sequence has certain restrictions associated with complementarity of the bases along DNA chain. These restrictions at the level of triplet sequence can be eliminated by degeneracy of the genetic code.

Keywords: genome, nucleotide sequence, Markov's chains, Bayesian information criterion, sequencing.

ВВЕДЕНИЕ

Геном – уникальная структура организма, в которой заключена огромная информация о строении организма, его функционировании, репродукции и т.д. В основе генома лежит материальная структура – молекула ДНК (дезоксирибонуклеиновой кислоты).

Ген — это участок молекулы ДНК, кодирующий первичную структуру молекулы белка, а также несущий другую важную информацию, необходимую для жизнедеятельности организма. Важнейшей характеристикой ДНК является ее нуклеотидный состав. ДНК впервые была выделена и изучена Фридрихом Мишером, который начал свои исследования в 1869 г. ДНК построена из четырех мононуклеотидных единиц: дАМФ (дезоксиаденозин-5'-фосфорная кислота), дГМФ (дезоксигуанозин-5'-фосфорная кислота), дТМФ (дезокситимидин-5'-фосфорная кислота), дЦМФ (дезоксицитидин-5'-фосфорная кислота) или, сокращенно, из четырех оснований (нуклеотидов) A (аденин), G (гуанин), T (тимин), C (цитозин) [1].

ЦЕЛЬ РАБОТЫ

Биофизический анализ некоторых принципов хранения информации с помощью цепи ДНК.

ПРИНЦИПЫ ПОСТРОЕНИЯ МОЛЕКУЛЫ ДНК

Частота встречаемости разных оснований в молекулах ДНК разных организмов различная. Но она не несет генетической информации.

Согласно модели Уотсона и Крика, молекула ДНК состоит из двух полимерных цепочек. Последовательность нуклеотидов в каждой цепи может быть произвольной, но она ограничена принципом их комплементарности:

— против A должно быть T, и наоборот;

— против G должно быть C, и наоборот.

Молекулы ДНК состоят из динуклеотидов *AT, TA, GC, CG* и имеют очень большую длину — до 2 м. Число возможных перестановок в последовательности n нуклеотидов в каждой полимерной цепи равно 4^n . С помощью данного четырехбуквенного алфавита можно закодировать практически бесконечное число генетической информации. Если принять молекулярный вес гена, состоящего из определенной последовательности динуклеотидов, $\sim 10^6$, то возможное число таких генов равно $\sim 4^{1500}$ [2]. Это число значительно превышает число всех генов, существовавших с момента зарождения жизни. Обычно ген содержит от 900 до 1500 динуклеотидов.

На основе информации, заключенной в ДНК, синтезируются белки. Этот процесс идет в два этапа.

На первом этапе специальный фермент РНК-полимераза снимает с гена (кодирующего белок участка ДНК) копию (матрицу) в виде молекулы мРНК (матричной рибонуклеиновой кислоты). В отличие от ДНК молекула мРНК состоит из одной полимерной цепочки. Этот этап получил название транскрипция (перезапись). При этом по какой-то причине тимин T заменен на урацил U, что не мешает процессу транскрипции. Урацил U также является комплементарным аденину A.

На втором этапе с помощью органоида клетки рибосомы синтезируется белок. Синтез белка осуществляется на основе генетического года. Этот этап называется трансляция (перевод).

Впервые правильный принцип построения генетического кода указал Г.А. Гамов [3]. Происхождение генетического кода проанализировано в [4].

ОСОБЕННОСТИ ФОРМИРОВАНИЯ ГЕНЕТИЧЕСКОГО КОДА

Генетический код представляет собой соответствие между последовательностью нуклеотидов в мРНК и аминокислотной последовательностью белка. Одной аминокислоте соответствует последовательность из трех нуклеотидов в мРНК. Последовательность трех нуклеотидов называется кодон. Количество аминокислотных остатков всех белков организма равно 20. Количество кодонов равно $4^3=64$. Поэтому часть аминокислот кодируется несколькими кодонами — от 1 до 6. Генетический код выглядит следующим образом [5], **таблица 1**.

В колонке 2 показаны возможные аминокислотные остатки, в колонке 3 — их сокращенные обозначения.

N	1	2	3	4	5	6	7	8	9
1	Met	M	1	ATG					
2	Trp	W	1	TGG					
3	Phe	F	2	TTT	TTC				
4	Tyr	Y	2	TAT	TAC				
5	His	H	2	CAT	CAC				
6	Asn	N	2	AAT	AAC				
7	Asp	D	2	GAT	GAC				
8	Cys	C	2	TGT	TGC				
9	Gln	Q	2	CAA	CAG				
10	Lys	K	2	AAA	AAG				
11	Glu	E	2	GAA	GAG				
12	Ile	I	3	ATA	ATT	ATC			
13	Val	V	4	GTT	GTC	GTA	GTG		
14	Pro	P	4	CCT	CCC	CCA	CCG		
15	Thr	T	4	ACT	ACC	ACA	ACG		
16	Ala	A	4	GCT	GCC	GCA	GCG		
17	Gly	G	4	GGT	GGC	GGA	GGG		
18	Ser	S	6	TCT	TCC	TCA	TCG	AGT	AGC
19	Leu	L	6	CTT	CTC	CTA	CTG	TTA	TTG
20	Arg	R	6	CGT	CGC	CGA	CGG	AGA	AGG
	ter	*	3	TAA	TAG	TGA			

Таблица 1. *Генетический код*

В колонке 4 указано, сколькими кодонами кодируется аминокислотный остаток (так называемая степень вырождения кодировки). В колонках 4—9 показаны кодоны, кодирующие аминокислотные остатки.

Например:

— аминокислотный остаток *Trp* кодируется одним кодоном TGG;

— аминокислотный остаток Asn кодируется двумя кодонами AAT и AAC;

— аминокислотный остаток *Arg* кодируется шестью кодонами CGE, CGC, CGA, CGG, AGA и AGG.

В конце **таблицы 1** показаны кодоны терминации (конца синтеза данного белка). Каждый из этих кодонов останавливает синтез белка.

Генетический код в целом универсален для всей живой природы. Однако существуют и исключения. Код митохондрий несколько отличается от основного кода. Это указывает на то, что генетический код является результатом эволюции.

ПРИНЦИПЫ АНАЛИЗА НУКЛЕОТИДНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ ДНК ЦЕПИ МАРКОВА

Анализ нуклеотидных последовательностей осуществляется различными способами. Часто необходимо провести сравнение двух участков нуклеотидных последовательностей. При этом используют так называемые точечные диаграммы [6]. По осям координат записывают исследуемые последовательности. В местах пересечения, соответствующих одному и тому же нуклеотиду, ставят точку. Далее исследуют количество точек на главной диагонали.

	TG	CT	CC	AG	AA	CA	GG	TT	GA	TC	GC	AT	AC	GT	TA	CG
$\frac{p_{uv}}{p_u p_v} < 1$	1,29	1,26	1,18	1,16	1,15	1,15	1,14	1,07	1,04	1,00	0,99	0,85	0,84	0,82	0,65	0,42

Таблица 2. *Комплементарность оснований вдоль молекулы ДНК*

Мы рассмотрим принципы анализа нуклеотидных последовательностей в цепи ДНК с использованием теории цепей Маркова [7].

Встречаемость разных оснований в молекуле ДНК неодинакова. Частоты пар соседних вдоль цепи ДНК оснований p_{uv} отличаются от произведений частот самих оснований $p_{uv} \neq p_u p_v$, где u и v – типы оснований. Это указывает на зависимость вероятностей встречаемости оснований в паре вдоль молекулы ДНК друг от друга. Вдоль молекулы ДНК положение основания определяется не только синтезируемым белком.

В **таблице 2** показаны некоторые относительные частоты динуклеотидов вдоль ДНК позвоночных. Не все основания полностью комплементарны (совместимы) друг другу. При $\frac{p_{uv}}{p_u p_v} < 1$ комплементарность оснований вдоль цепи ДНК снижена. На нуклеотидную последовательность кодирующих участков налагаются строгие ограничения, связанные с последовательностью аминокислотных остатков синтезируемых белков. На эти ограничения накладываются ограничения на уровне кодонов, связанные с ограничениями на динуклеотидном (вдоль ДНК) уровне. Ограничения на уровне кодонов могут в значительной мере сниматься вырожденностью генетического кода. По-видимому, также ограничения на уровне кодонов снижают мутационную устойчивость цепи ДНК, способствуя эволюционным процессам.

Марковская цепь определяется как последовательность случайных величин, обладающая тем свойством, что распределение величины x_n зависит только от значения x_{n-1} . Последовательность нуклеотидов в цепи ДНК можно считать марковской цепью. В цепи Маркова взаимозависимыми являются и удаленные друг от друга основания. Часто бывает необходимо определить расстояние, на которое распространяется взаимодействие между основаниями вдоль ДНК, выявить особенности последовательностей нуклеотидов, повторы или сходные участки последовательностей, мутационные замены отдельных нуклеотидов и т.д.

Для примера проведем первоначальное исследование марковской цепи из 9 нуклеотидов CTATAATAG.

Найдем вероятность, что за кодоном ATA следует нуклеотид A. Эта вероятность равна

$P(A \mid ATA) = \frac{n_{ATAA}}{n_{ATA}} = \frac{1}{2}$, где $n_{ATA} = 2$ — число кодонов ATA в последовательности, n_{ATTA}=1— число последовательностей ATAA в общей последовательности нуклеотидов.

Для дальнейшего анализа используем метод функции правдоподобия, предложенный выдающимся генетиком Р.А. Фишером. Метод функции правдоподобия, в частности, позволяет установить порядок цепи Маркова, то есть определить расстояние, на которое распространяется взаимодействие нуклеотидов.

Предположим, что исследуемая последовательность нуклеотидов узнается некоторым ферментом. Функция правдоподобия характеризует вероятность появления данной последовательности в общей последовательности цепи нуклеотидов. Но эта вероятность зависит от порядка марковской цепи последовательности нуклеотидов.

Последовательность, составленная из независимых оснований, будет соответствовать Марковской цепи 0-го порядка. Функция правдоподобия цепи Маркова нулевого порядка исследуемой последовательности нуклеотидов равна $L(0) = \frac{1}{4^9} = \frac{1}{262144}$.

Цепь порядка 1 предполагает, что вероятность нахождения какого-либо основания в позиции i зависит только от вероятности присутствия одного из четырех оснований в позиции i-1.

Функция правдоподобия, например, для марковской цепи нуклеотидов первого порядка вычисляется по формуле:

$$L(1) = P(C)P(T \mid C)P(A \mid T)P(T \mid A)P(A \mid T)P(A \mid A)P(T \mid A)P(A \mid T)P(G \mid A) = \frac{1}{4} \cdot \frac{n_{CT}}{n_C} \cdot \frac{n_{TA}}{n_T} \cdot \frac{n_{AT}}{n_A} \cdot \frac{n_{TA}}{n_T} \cdot \frac{n_{AA}}{n_A} \cdot \frac{n_{AT}}{n_A} \cdot \frac{n_{TA}}{n_T} \cdot \frac{n_{AG}}{n_A} = \frac{1}{4} \cdot 1 \cdot \frac{3}{3} \cdot \frac{2}{4} \cdot \frac{3}{3} \cdot \frac{1}{4} \cdot \frac{2}{4} \cdot \frac{3}{3} \cdot \frac{1}{4} = \frac{1}{256}.$$

Функция правдоподобия для марковской цепи нуклеотидов второго порядка, используя $P(CT) = P(C)P(T \mid C) = \frac{1}{4} \cdot 1 = \frac{1}{4}$, вычисляется аналогично: $L(2) = \frac{1}{108}$. Функция правдоподобия для цепи третьего порядка равна $L(3) = \frac{1}{16}$ и т.д.

Любая цепь, или последовательность, характеризуется своими параметрами. Например, последовательность единиц (или букв A) характеризуется одним параметром – единицей (или буквой A). Последовательность случайных чисел характеризуется тремя параметрами: математическим ожиданием, дисперсией и корреляционной (или ковариационной) функцией элементов последовательности. В связи с этим заметим, что справедлива эргодическая гипотеза, которая, в частности, предполагает, что математическое ожидание элемента последовательности во времени равно математическому ожиданию всех элементов последовательности в данный момент времени по длине цепи.

Цепь Маркова k-го порядка характеризуется $\varepsilon = 3 \cdot 4^k$ параметрами. Второй сомножитель определяет число возможных перестановок 4 нуклеотидов. Таким образом, для цепи Маркова 0-го порядка (последовательность независимых оснований или случайных чисел) число параметров равно 3, для марковской цепи первого порядка число параметров равно 12, второго порядка —

48, а третьего порядка — 192 параметра. Отношение функций правдоподобия следующих друг за другом порядков цепи Маркова обозначим $\lambda_k = \frac{L(k)}{L(k+1)}$.

Величина $-2\ln\frac{L(k)}{L(k+1)} = -2(\ln L(k) - \ln(k+1))$ подчиняется распределению χ^2 [7] с числом степеней свободы, равным разности параметров цепей Маркова. Для рассматриваемой модельной последовательности нуклеотидов варианты распределения χ^2 равны:
$-2\ln\frac{L(0)}{L(1)} = -2\ln\frac{108}{262144} = 15{,}6$ с числом степеней свободы $\nu = 12\text{-}3 = 9$, $-2\ln\frac{L(1)}{L(2)} = 1{,}73$ с числом степеней свободы $\nu = 48\text{-}12 = 46$, и $-2\ln\frac{L(2)}{L(3)} = 3{,}82$ с числом степеней свободы $\nu = 144$.

Для того чтобы выбрать адекватный порядок цепи нуклеотидной последовательности, отражающий истинный уровень связи нуклеотидов вдоль цепи ДНК (фактически, противодействующий кодированию белков), часто используется информационный критерий Байеса (Bayesian Information Criterion):

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

где n_k — число подпоследовательностей длины $k+1$, находящихся в рассматриваемой последовательности; n_k равно числу элементов последовательности минус k. То значение k, для которого $BIC(k)$ минимально, принимается за оценку. Постоянная величина роли не играет, так как осуществляется только сравнение критериев BIK(k). Ее можно условно принять равной нулю. Таким образом: $BIC(0) = Const - 2\ln L(0) + 3\ln(9-0) = 31{,}54$; $BIC(1) = 36{,}04$; $BIC(2) = 102{,}8$; $BIC(3) = 349{,}6$.

Естественно, расчет такой короткой последовательности носит в основном иллюстративный характер. Но уже этот расчет показывает, что нужно выбирать первый порядок цепи Маркова для адекватного анализа последовательности нуклеотидов. Нулевой порядок цепи неприемлем, поскольку нуклеотиды в ДНК нельзя считать независимыми.

В **таблице 3** представлены результаты расчета цепи Маркова для последовательности 431 нуклеотида мРНК кодирующей части генома β-глобина цыпленка [7].

Наименьший BIC получен для цепи первого порядка. Таким образом, цепь Маркова первого порядка лучше всего соответствует последовательности нуклеотидов кодирующей части генома β-глобина цыпленка.

k	BIC(k)	hL(k)	ε	χ^2	ν
0	1213,04	-597,49	3		
1	1196,19	-562,93	12	68,94	9
2	1365,7	-540,12	48	45,62	36
3	2068,46	-454,77	192	170,70	144
4	5334,38	-290,52	768	328,50	576

Таблица 3. *Параметры цепи Маркова для части генома цыпленка*

СЕКВЕНИРОВАНИЕ ЦЕПЕЙ ДНК

В заключение рассмотрим принцип секвенирования цепей ДНК.

Секвенирование – это определение нуклеотидной последовательности цепи ДНК.

Существует ряд методов, с помощью которых проводится секвенирование. Один из наиболее популярных методов, которым в частности впервые была установлена последовательность нуклеотидов в ДНК человека (проект «Геном человека» 1990—2003 г.), является метод секвенирования по Ф. Сэнгеру, за что ему была присуждена Нобелевская премия в 1980 г.

Упрощенно суть метода состоит в следующем. Выделенные из клеток молекулы ДНК делят на рестрикционные фрагмены с помощью рестриктаз – ферментов, которые разрезают молекулу ДНК, после распознавания ими определенной нуклеотидной последовательности. Например, рестриктаза BamH1 распознает последовательность GGATCC, а рестриктаза EcoRI — последовательность GAATTC.

Затем осуществляется процесс амплификации фрагментов ДНК. Амплификация – это накопление копий фрагментов ДНК. Амплификация осуществляется следующим образом. Фрагменты ДНК сначала денатурируют путем их нагревания, получая отдельные нити. Затем добавляют так называемые праймеры. Праймеры – это короткие фрагменты нуклеотидной последовательности (несколько нуклеотидов), соединяющиеся с фрагментом нити ДНК и служащие для начала работы ДНК-полимеразы.

Далее полученную смесь разделяют на четыре части, в каждую из которых добавляют один из дидезоксинуклеотидов (далее dd): ddATP, ddCTP, ddGTP или ddTTP и ДНК-полимеразу. ДНК-полимераза катализирует полимеризацию нити ДНК, комплементарную копируемому фрагменту, с помощью полимеразной цепной реакции. В результате получается копия фрагмента. Но эта копия заканчивается, если ДНК-полимераза захватывает dd.

В результате получаются фрагменты разной длины, но с определенным dd на конце. Участок фрагмента перед dd называется префиксом.

Допустим, мы определяем последовательность в цепи 9 нуклеотидов CTATAATAG. В 4 пробирках образуются префиксы C, CT, CTA, CTAT, CTATA, CTATAA, CTATAAT, CTATAATA, CTATAATAG. Если полученные смеси подвергнуть электрофорезу в геле на 4 дорожках, то более длинная последовательность пройдет более короткий путь и последовательности разделятся по длине. Выяснив каким-либо методом (например, методом радиоакивно меченного дезоксинуклеотида), какой нуклеотид в самом коротком префиксе (C), мы находим начальную букву последовательности. Затем мы выясняем, какие нуклеотиды в более длинном префиксе (CT), и находим вторую букву последовательности, и т.д.

В настоящее время методы секвенирования в значительной мере автоматизированы и процесс при относительно низкой стоимости не занимает много времени.

ЗАКЛЮЧЕНИЕ

Последовательность нуклеотидов в цепи ДНК является основой хранения информации о строении организма, его функционировании, репродукции и т.д. Эта информация записывается с помощью генетического кода. Однако на последовательность нуклеотидов накладываются ограничения, связанные с комплементарностью оснований вдоль цепи ДНК. Эти ограничения на уровне последовательности кодонов могут в значительной мере сниматься вырожденностью генетического кода. По-видимому, также ограничения на уровне кодонов снижают мутационную устойчивость цепи ДНК, способствуя эволюционным процессам.

С целью сравнения последовательности нуклеотидов в ДНК, определения расстояния, на которое распространяется взаимодействие между основаниями вдоль ДНК, выявления особенностей последовательностей нуклеотидов, повторов или сходных участков последовательностей, мутационных замен отдельных нуклеотидов и т.д. проводится генетико-математический анализ нуклеотидных последовательностей. Использование теории цепей Маркова позволяет определить истинный уровень связи нуклеотидов вдоль цепи ДНК.

Секвенирование нуклеотидной последовательности в настоящее время в значительной мере автоматизировано, имеет относительно низкую стоимость и не занимает много времени.

ЛИТЕРАТУРА / REFERENCES

1. Ленинджер А. Биохимия. Молекулярные основы структуры и функций клетки. М.: Мир, 1974. [Lehninger A. Biokhimiya. Molekulyarnye osnovy struktury i funktsii kletki. (In Russ.)].
2. Уотсон Дж. Молекулярная биология гена. М.: Мир, 1978 [Watson JD. Molekulyarnaya biologiya gena. (In Russ.)].
3. Gamov GA. Possible Relation between Deoxyribonucleic Acid and Protein Structures. Nature 173 Feb 13(1954):318.
4. Резерфорд А. Биография жизни. От первой клетки до генной инженерии. М.: БИНОМ. Лаборатория знаний, 2016 [Rutherford A.Biografiya zhizni. Ot pervoi kletki do gennoi inzhenerii, 2013. (In Russ.)].
5. Козлов Н.Н. Математический анализ генетического кода. М.: БИНОМ, 2015 [Kozlov NN. Matematicheskij analiz geneticheskogo koda. M.: BINOM, 2015. (In Russ.)].
6. Brown D., Rothery P. Models in Biology: Mathematics, Statistics, and Computing. Jon Wiley & Sons Ltd., Chichester, NY, 1994
7. Вейр Б. Анализ генетических данных. М: Мир, 1995. [Weir B.S. Analiz geneticheskikh dannykh. (In Russ.)].

Участие авторов

Разработка концепции исследования, математический анализ и расчеты, написание текста статьи: Волобуев А.Н.
Участие в разработке концепции исследования: Петров Е.С.
Редактирование текста статьи: Романчук Н.П.

Конфликт интересов отсутствует.

СВЕДЕНИЯ ОБ АВТОРАХ

Волобуев А.Н. — д.т.н., профессор, заведующий кафедрой физики, математики и информатики Самарского государственного медицинского университета.
E-mail: volobuev47@yandex.ru

Петров Е.С. — к.м.н., доцент, доцент кафедры оперативной хирургии и топографической анатомии с курсом инновационных технологий Самарского государственного медицинского университета.
E-mail: petroves@inbox.ru

Романчук Н.П. — аспирант физиологии с курсом безопасности жизнедеятельности и медицины катастроф Самарского государственного медицинского университета.
E-mail: romanchuknp@mail.ru

INFORMATION ABOUT AUTHORS

Volobuev AN – PhD, Professor,
head of the Department of medical physics,
mathematics and informatics,
Samara State Medical University.
E-mail: volobuev47@yandex.ru

Petrov ES — PhD, Associate Professor, associate professor of the Department of operative surgery and topographical anatomy with the course of innovative technologies, Samara State Medical University.
E-mail: petroves@inbox.ru

Romanchuk NP — post-graduate student of the Department of physiology with the course of health and safety and disaster medicine, Samara State Medical University.
E-mail: romanchuknp@mail.ru

Контактная информация

Волобуев Андрей Николаевич
Адрес: а/я 1423, г. Самара, Россия, 443079.
Телефон: +7 (927) 016 46 95
E-mail: volobuev47@yandex.ru

Contact information

Volobuev Andrei Nikolaevich
Address: postbox 1423, Samara, Russia, 443079.
Phone: +7 (927) 016 46 95
E-mail: volobuev47@yandex.ru