УДК 519.217.2

И.В. СЕРГИЕНКО, А.М. ГУПАЛ, А.А. ВАГИС

ПРАВИЛА СИММЕТРИИ В ЗАПИСИ ГЕНЕТИЧЕСКОЙ ИНФОРМАЦИИ В ДНК[1]

Ключевые слова: *основания, комплементарность, симметрия, цепь Маркова, переходные вероятности.*

ВВЕДЕНИЕ

Соотношения комплементарности, или симметрии, в записи оснований, подсчитанных по одной нити в хромосомах ДНК, исследовались в работах [1–3] (в [1] содержится список литературы по данному вопросу). В [2, 3] соотношения симметрии приведены в виде коротких формул, что значительно упрощает восприятие этих результатов и является основой построения математического аппарата в целях получения новых результатов. Статистический анализ подтвердил выполнение соотношений симметрии относительно геномов бактерий, растений, высших организмов (примерно сто геномов), в том числе в ДНК человека [2, 3]. Таким образом, в записи генетической информации в ДНК явно наблюдается симметрия, однако до настоящего времени не выяснены причины, которые объясняют этот феномен в природе.

В настоящей работе получены новые правила в записи оснований по одной нити в хромосомах ДНК. Доказано, что из симметрии последовательностей оснований вытекает симметрия коротких последовательностей, в том числе отдельных оснований. Для пар оснований возможны два вида симметрии, но в природе реализован один, более эффективный способ записи и считывания информации.

На основе модели однородной цепи Маркова показано, что симметрия для троек, четверок и коротких последовательностей оснований вытекает из симметрии пар оснований.

СИММЕТРИЯ ОСНОВАНИЙ

ДНК имеет форму двойной спирали, информация записана в четырехбуквенном алфавите оснований: аденин (A), цитозин (C), гуанин (G), тимин (T). Известно, что C — G, A — T — комплементарные пары оснований, связывающие две цепи. Хромосомы — неделимые участки ДНК, в них содержится информация относительно тысяч генов, поэтому расчеты проводились на уровне всей хромосомы, а не отдельного гена.

Запись и считывание оснований по первой комплементарной нити хромосомы ДНК выполняется слева направо в направлении $5' \to 3'$, по второй — справа налево в направлении $5' \to 3'$ (рис. 1). Приводимые далее соотношения, как правило, выполняются приближенно.

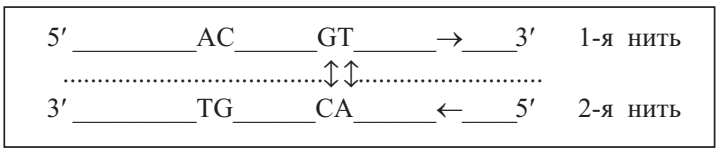

Рис 1. Условная запись двух нитей хромосомы

[1]Работа выполнена в рамках проекта НАН Украины и Российского фонда фундаментальных исследований 2010–2011 гг. при финансовой поддержке Президиума НАН Украины.

Для оснований, записанных по одной нити ДНК хромосомы, выполняются приближенные соотношения

$$n(\mathrm{A}) = n(\mathrm{T}),\ \ n(\mathrm{C}) = n(\mathrm{G})\,, \tag{1}$$

где $n(j)$ — количество оснований j, $j \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$, вычисленных на одной нити.

Заметим, что из комплементарности пар оснований по двум нитям ДНК не следует, что значения букв A и T, а также C и G, подсчитанные по одной нити, совпадают между собой.

Из соотношений (1) вытекает, что количества каждого основания, подсчитанного по первой и второй нити, совпадают:

$$n(\mathrm{A},1) = n(\mathrm{A},2),\ \ n(\mathrm{T},1) = n(\mathrm{T},2),\quad n(\mathrm{C},1) = n(\mathrm{C},2),\ \ n(\mathrm{G},1) = n(\mathrm{G},2)\,. \tag{2}$$

Таким образом, имеет место симметрия относительно записи оснований по каждой нити ДНК. Отсюда следует важный вывод о том, что веса двух нитей совпадают.

СИММЕТРИЯ ПАР ОСНОВАНИЙ

Расчеты показали, что для пар оснований выполняются соотношения

$$\begin{gathered} n(\mathrm{AC}) = n(\mathrm{GT}),\, n(\mathrm{AG}) = n(\mathrm{CT}), \\ n(\mathrm{TC}) = n(\mathrm{GA}),\, n(\mathrm{TG}) = n(\mathrm{CA}), \\ n(\mathrm{AA}) = n(\mathrm{TT}),\, n(\mathrm{CC}) = n(\mathrm{GG}), \end{gathered} \tag{3}$$

или короче, в виде формулы

$$n(ij) = n(\bar{j}\bar{i}), \tag{4}$$

где $i, j \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$, $\overline{\mathrm{A}} = \mathrm{T}$, $\overline{\mathrm{C}} = \mathrm{G}$, $\overline{\mathrm{T}} = \mathrm{A}$, $\overline{\mathrm{G}} = \mathrm{C}$. Заметим, что пары AT, TA, CG и GC не входят в (3), поскольку они приводят к тавтологии. В табл. 1 приведены значения пар оснований в геноме человека [2].

Таблица 1

Пара оснований	Количество пар оснований в геноме человека по одной нити ДНК				
	Хромосома 1	Хромосома 3	Хромосома 6	Хромосома 10	Хромосома 18
AA	21 191 409	19 746 023	17 083 089	12 607 303	7 553 856
TT	21 245 312	19 772 366	17 080 492	12 628 305	7 560 778
AC	11 189 673	9 791 735	8 417 550	6 641 892	3 762 190
GT	11 209 763	9 798 222	8 411 037	6 651 425	3 776 890
AG	15 878 823	13 482 539	11 543 173	9 275 834	5 136 579
CT	15 904 404	13 478 613	11 532 563	9 286 062	5 138 944
CA	16 200 299	13 972 734	11 983 646	9 656 789	5 382 301
TG	16 226 750	13 970 283	11 984 196	9 667 666	5 401993
CC	12 132 633	9 518 322	8 128 472	7 073 095	3 640 163
GG	12 121 539	9 520 091	8 140 958	7 062 604	3 647 384
GA	13 313 713	11 472 583	9 879 809	7 851 856	4 411 285
TC	13 322 934	11 477 596	9 862 177	7 860 740	4 408 666
AT	16 615 348	15 646 889	13 495 077	9 896 788	6 012 563
TA	14 169 829	13 466 193	11 592 344	8 305 870	5 117 737
CG	2 256 627	1 620 941	1 473 327	1 353 534	677 210
GC	9 838 754	7 836 943	6 709 818	5 793 769	3 027 601

Из соотношений (3), (4) вытекает симметрия относительно записи 16 пар оснований по каждой нити ДНК

$$n(ij,1) = n(ij,2), \tag{5}$$

где $i, j \in \{\mathrm{A},\mathrm{C},\mathrm{G},\mathrm{T}\}$.

Известно, что соотношения

$$\hat{p}(ij) = \frac{n(ij)}{n(i)}, \tag{6}$$

где $n(ij)$ — число пар (ij), $i, j \in \{\mathrm{A},\mathrm{C},\mathrm{G},\mathrm{T}\}$, $n(i)$ — число оснований i в цепи хромосомы, представляют собой оценки переходных вероятностей для однородных цепей Маркова [4].

В отличие от независимых бернуллиевских величин математическое ожидание оценок переходных вероятностей, построенных в виде частот, смещено и не совпадает с точными значениями вероятностей. В [4] показано, что оценки переходных вероятностей асимптотически нормальны, и выведены формулы дисперсии и ковариации оценок для этого предельного распределения.

Из (5) и (6) вытекает, что вторая комплементарная нить в направлении $5' \to 3'$ имеет такие же оценки переходных вероятностей $\hat{p}(ij)$, как и исходная первая нить (рис. 1). Отсюда следует, что вероятности двух противоположных нитей хромосомы, подсчитанные в модели однородной цепи Маркова на основе оценок переходных вероятностей (6), совпадают.

Легко заметить, что для любой последовательности без пропусков букв с точностью до единицы выполняются соотношения

$$n(i) = n(\mathrm{A}i) + n(\mathrm{C}i) + n(\mathrm{G}i) + n(\mathrm{T}i) =$$
$$= n(i\mathrm{A}) + n(i\mathrm{C}) + n(i\mathrm{G}) + n(i\mathrm{T}), \tag{7}$$

где $i \in \{\mathrm{A},\mathrm{C},\mathrm{G},\mathrm{T}\}$, т.е. количество каждой буквы текста можно подсчитать на основе количеств пар букв.

Для основания A из (7) получаем связывающее ограничение для пар AT, TA, которые не входят в (3),

$$n(\mathrm{CA}) + n(\mathrm{GA}) + n(\mathrm{TA}) = n(\mathrm{AC}) + n(\mathrm{AG}) + n(\mathrm{AT}), \tag{8}$$

для основания C из (7) — ограничение для пар CG и GC

$$n(\mathrm{AC}) + n(\mathrm{GC}) + n(\mathrm{TC}) = n(\mathrm{CA}) + n(\mathrm{CG}) + n(\mathrm{CT}). \tag{9}$$

Для оснований T и G с учетом (3) получаем те же соотношения, что и в (8), (9).

Например, для хромосомы 6 генома человека (табл. 1) имеем

$$n(\mathrm{CA}) + n(\mathrm{GA}) + n(\mathrm{TA}) = 33\,455\,799, \quad n(\mathrm{AC}) + n(\mathrm{AG}) + n(\mathrm{AT}) = 33\,455\,800.$$

Одна из особенностей в анализе последовательностей оснований состоит в том, что частоты встречаемости соседних букв не являются независимыми [5]. Частоты пар соседних оснований отличаются от произведений частот самих оснований, т.е. $\frac{n(ij)}{n} \neq \frac{n(i)}{n}\frac{n(j)}{n}$, $i, j \in \{\mathrm{A},\mathrm{C},\mathrm{G},\mathrm{T}\}$, n — длина хромосомы. Это очевидно, например, для пар CG и GC в табл. 1.

Заметим, что симметрия (5) может наблюдаться также и для соотношений $n(ij) = n(\bar{i}\bar{j})$, т.е. когда обе нити хромосомы имеют одинаковые направления записи и считывания оснований. Однако в природе такой вид симметрии не реализован и в данной работе не рассматривается, поскольку требует отдельного обсуждения. У симметрии вида $n(ij) = n(\bar{j}\bar{i})$ больше степеней свободы, чем у симмет-

рии $n(ij) = n(\bar{i}\bar{j})$, и с точки зрения теории информации она более эффективна. Поэтому из симметрии оснований (1), (2) нельзя вывести симметрию пар оснований. Обратное утверждение справедливо.

Утверждение 1. Из симметрии пар оснований вытекает симметрия оснований.

Действительно, с помощью соотношений (7) количества оснований $n(\mathrm{A})$ и $n(\mathrm{T})$ записываются в виде равенств

$$n(\mathrm{A}) = n(\mathrm{AA}) + n(\mathrm{AC}) + n(\mathrm{AG}) + n(\mathrm{AT}),$$

$$n(\mathrm{T}) = n(\mathrm{TA}) + n(\mathrm{TC}) + n(\mathrm{TG}) + n(\mathrm{TT}).$$

С учетом $n(ij) = n(\bar{j}\bar{i})$, $i, j \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$, и соотношения (8), которое принимает вид

$$n(\mathrm{AC}) + n(\mathrm{AG}) + n(\mathrm{AT}) = n(\mathrm{TA}) + n(\mathrm{GA}) + n(\mathrm{CA}),$$

получаем $n(\mathrm{A}) = n(\mathrm{T})$.

Вывод равенства $n(\mathrm{C}) = n(\mathrm{G})$ выполняется аналогично на основе (7), (9) и соотношений $n(ij) = n(\bar{j}\bar{i})$, $i, j \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$:

$$n(\mathrm{AC}) + n(\mathrm{GC}) + n(\mathrm{TC}) = n(\mathrm{CA}) + n(\mathrm{CG}) + n(\mathrm{CT}).$$

СИММЕТРИЯ ТРОЕК ОСНОВАНИЙ

Кодоны (тройки оснований) связаны следующими соотношениями:

$$n(ijk) = n(\bar{k}\bar{j}\bar{i}). \tag{10}$$

Здесь $n(ijk)$ — число троек оснований (ijk), $i, j, k \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$, $(\bar{k}\bar{j}\bar{i})$ — антикодон кодона (ijk). В работе [2] для 64 триплетов получены 32 соотношения вида (10) типа кодон–антикодон в хромосоме 6 генома человека (табл. 2).

Таблица 2

Кодон	Число кодонов	Кодон	Число кодонов	Кодон	Число кодонов	Кодон	Число кодонов
AAA	6 742 017	TTT	6 744 661	CAG	3 216 761	CTG	3 217 346
AAC	2 509 339	GTT	2 507 886	CCA	2 932 409	TGG	2 932 367
AAG	3 412 535	CTT	3 407 422	CCC	1 980 135	GGG	1 986 846
AAT	4 419 198	ATT	4 420 523	CCG	394 680	CGG	396 760
ACA	3 417 383	TGT	3 417 331	CGA	341 096	TCG	340 572
ACC	1 872 766	GGT	1 869 465	CGC	345 302	GCG	346 653
ACG	391 422	CGT	390 169	CTA	2 226 977	TAG	2 227 635
ACT	2 735 979	AGT	2 734 072	CTC	2 680 818	GAG	2 686 241
AGA	3 741 389	TCT	3 735 896	GAA	3 394 901	TTC	3 388 807
AGC	2 242 727	GCT	2 239 440	GAC	1 533 503	GTC	1 532 047
AGG	2 824 985	CCT	2 821 248	GCA	2 330 699	TGC	2 327 157
ATA	3 684 661	TAT	3 682 369	GCC	1 793 026	GGC	1 794 632
ATC	2 260 505	GAT	2 265 164	GGA	2 490 014	TCC	2 482 545
ATG	3 129 388	CAT	3 128 346	GTA	1 962 626	TAC	1 966 011
CAA	3 229 842	TTG	3 228 944	TAA	3 716 329	TTA	3 718 080
CAC	2 408 697	GTG	2 408 478	TCA	3 303 155	TGA	3 307 301

Аналогично (5) из соотношений (10) вытекает симметрия относительно записи 64 троек оснований для каждой нити ДНК

$$n(ijk, 1) = n(ijk, 2). \tag{11}$$

Для 16 пар оснований (ij), $i, j \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$, справедливы соотношения

$$n(ij) = n(\mathrm{A}ij) + n(\mathrm{C}ij) + n(\mathrm{G}ij) + n(\mathrm{T}ij) = n(ij\mathrm{A}) + n(ij\mathrm{C}) + n(ij\mathrm{G}) + n(ij\mathrm{T}).$$

Например, из табл. 2 имеем

$$n(\mathrm{AAA}) + n(\mathrm{AAC}) + n(\mathrm{AAG}) + n(\mathrm{AAT}) = 17\,083\,089,$$

$$n(\mathrm{AAA}) + n(\mathrm{CAA}) + n(\mathrm{GAA}) + n(\mathrm{TAA}) = 17\,083\,089.$$

Для шести пар оснований (3), используя соотношения (10), получаем следующие связывающие ограничения:

$$n(\mathrm{AAC}) + n(\mathrm{AAG}) + n(\mathrm{AAT}) = n(\mathrm{CAA}) + n(\mathrm{GAA}) + n(\mathrm{TAA}), \tag{12}$$

$$n(\mathrm{ACA}) + n(\mathrm{ACC}) + n(\mathrm{ACG}) + n(\mathrm{ACT}) = n(\mathrm{AAC}) + n(\mathrm{CAC}) + n(\mathrm{GAC}) + n(\mathrm{TAC}), \tag{13}$$

$$n(\mathrm{AGA}) + n(\mathrm{AGC}) + n(\mathrm{AGG}) + n(\mathrm{AGT}) =$$
$$= n(\mathrm{AAG}) + n(\mathrm{CAG}) + n(\mathrm{GAG}) + n(\mathrm{TAG}), \tag{14}$$

$$n(\mathrm{CAA}) + n(\mathrm{CAC}) + n(\mathrm{CAG}) + n(\mathrm{CAT}) = n(\mathrm{ACA}) + n(\mathrm{CCA}) + n(\mathrm{GCA}) + n(\mathrm{TCA}), \tag{15}$$

$$n(\mathrm{CCA}) + n(\mathrm{CCG}) + n(\mathrm{CCT}) = n(\mathrm{ACC}) + n(\mathrm{GCC}) + n(\mathrm{TCC}), \tag{16}$$

$$n(\mathrm{GAA}) + n(\mathrm{GAC}) + n(\mathrm{GAG}) + n(\mathrm{GAT}) = n(\mathrm{AGA}) + n(\mathrm{CGA}) + n(\mathrm{GGA}) + n(\mathrm{TGA}). \tag{17}$$

Для пар AT, TA, CG и GC новые соотношения не выводятся, поскольку из (10) получаем тавтологии. Формулы (12)–(17) важны тем, что с помощью универсального генетического кода они переводятся в соотношения для аминокислот.

Утверждение 2. Из симметрии троек оснований вытекает симметрия пар оснований.

С помощью равенств (7) для пар букв количества $n(\mathrm{AA})$ и $n(\mathrm{TT})$ записываются в виде соотношений:

$$n(\mathrm{AA}) = n(\mathrm{AAA}) + n(\mathrm{AAC}) + n(\mathrm{AAG}) + n(\mathrm{AAT}),$$

$$n(\mathrm{TT}) = n(\mathrm{TTA}) + n(\mathrm{TTC}) + n(\mathrm{TTG}) + n(\mathrm{TTT}).$$

Данные выражения с помощью $n(ijk) = n(\bar{k}\bar{j}\bar{i})$ преобразуются в равенство (12):

$$n(\mathrm{AAC}) + n(\mathrm{AAG}) + n(\mathrm{AAT}) = n(\mathrm{TAA}) + n(\mathrm{GAA}) + n(\mathrm{CAA}).$$

Доказательство для остальных пар $n(\mathrm{CC}) = n(\mathrm{GG})$, $n(\mathrm{AC}) = n(\mathrm{GT})$, $n(\mathrm{AG}) = n(\mathrm{CT})$, $n(\mathrm{CA}) = n(\mathrm{TG})$, $n(\mathrm{GA}) = n(\mathrm{TC})$ проводится аналогично с учетом формул (13)–(17).

Таким образом, из симметрии последовательностей оснований по индукции вытекает симметрия коротких последовательностей.

Заметим, что соотношения (12)–(17) выполняются для модели однородной цепи Маркова:

$$\hat{p}(\mathrm{AAC}) + \hat{p}(\mathrm{AAG}) + \hat{p}(\mathrm{AAT}) = \hat{p}(\mathrm{CAA}) + \hat{p}(\mathrm{GAA}) + \hat{p}(\mathrm{TAA}),$$

$$\hat{p}(\mathrm{AAC}) + \hat{p}(\mathrm{AAG}) + \hat{p}(\mathrm{AAT}) = \frac{n(\mathrm{AA})n(\mathrm{AC})}{nn(\mathrm{A})} + \frac{n(\mathrm{AA})n(\mathrm{AG})}{nn(\mathrm{A})} +$$

$$+ \frac{n(\mathrm{AA})n(\mathrm{AT})}{nn(\mathrm{A})} = \frac{n(\mathrm{AA})(n(\mathrm{AC}) + n(\mathrm{AG}) + n(\mathrm{AT}))}{nn(\mathrm{A})},$$

$$\hat{p}(\mathrm{CAA}) + \hat{p}(\mathrm{GAA}) + \hat{p}(\mathrm{TAA}) = \frac{n(\mathrm{CA})n(\mathrm{AA})}{nn(\mathrm{A})} + \frac{n(\mathrm{GA})n(\mathrm{AA})}{nn(\mathrm{A})} + \frac{n(\mathrm{TA})n(\mathrm{AA})}{nn(\mathrm{A})} =$$

$$= \frac{n(\mathrm{AA})(n(\mathrm{CA}) + n(\mathrm{GA}) + n(\mathrm{TA}))}{nn(\mathrm{A})}.$$

Остается воспользоваться формулой (8). Для обоснования (13), учитывая формулу (7), получаем

$$\hat{p}(\mathrm{ACA}) + \hat{p}(\mathrm{ACC}) + \hat{p}(\mathrm{ACG}) + \hat{p}(\mathrm{ACT}) =$$

$$= \frac{n(\mathrm{AC})(n(\mathrm{CA}) + n(\mathrm{CC}) + n(\mathrm{CG}) + n(\mathrm{CT}))}{nn(\mathrm{C})} = \frac{n(\mathrm{AC})n(\mathrm{C})}{nn(\mathrm{C})},$$

$$\hat{p}(\mathrm{AAC}) + \hat{p}(\mathrm{CAC}) + \hat{p}(\mathrm{GAC}) + \hat{p}(\mathrm{TAC}) =$$

$$= \frac{n(\mathrm{AC})(n(\mathrm{AA}) + n(\mathrm{CA}) + n(\mathrm{GA}) + n(\mathrm{TA}))}{nn(\mathrm{A})} = \frac{n(\mathrm{AC})n(\mathrm{A})}{nn(\mathrm{A})}.$$

Аналогичным образом выполняются формулы (14)–(17).

Поскольку симметрия в записи оснований по нитям в ДНК обнаружена эмпирически и в настоящее время не существует объяснения этого феномена в природе, важно построить модель, подтверждающую симметрию последовательностей оснований на основе симметрии коротких последовательностей.

Утверждение 3. Для модели однородной цепи Маркова симметрия троек оснований вытекает из симметрии оснований и симметрии пар оснований.

Из соотношений (1), (4) следует, что для однородной цепи Маркова оценки вероятностей троек оснований (ijk) и $(\bar{k}\bar{j}\bar{i})$ совпадают:

$$n\hat{p}(ijk) = \frac{n(i)n(ij)n(jk)}{n(i)n(j)} = n\hat{p}(\bar{k}\bar{j}\bar{i}) = \frac{n(\bar{k})n(\bar{k}\bar{j})n(\bar{j}\bar{i})}{n(\bar{k})n(\bar{j})},$$

где n — длина хромосомы. Таким образом, ожидаемое число повторов троек оснований (ijk) и $(\bar{k}\bar{j}\bar{i})$ совпадает по длине хромосомы.

Симметрия для последовательностей оснований также подтверждается для модели однородной цепи Маркова и вытекает из симметрии пар оснований. Этот результат является следствием важного утверждения.

Утверждение 4. Оценка вероятности последовательности $x_1, x_2, \ldots, x_{n-1}, x_n$ совпадает с оценкой вероятности последовательности $\bar{x}_n, \bar{x}_{n-1}, \ldots, \bar{x}_2, \bar{x}_1$, т.е.

$$\hat{p}(x_1, x_2, \ldots, x_{n-1}, x_n) = \hat{p}(\bar{x}_n, \bar{x}_{n-1}, \ldots, \bar{x}_2, \bar{x}_1). \tag{18}$$

Вероятность однородной цепи Маркова определяется соотношением

$$p(x_1, x_2, \ldots, x_{n-1}, x_n) = p(x_1)p(x_1, x_2)\ldots p(x_{n-1}, x_n), \tag{19}$$

где $p(x_1)$ — вероятность начального состояния, $p(x_{i-1}, x_i)$ — переходные вероятности, $i = 1, 2, \ldots, n$.

Заменив вероятность начального состояния частотой, а переходные вероятности $p(x_{i-1}, x_i)$ в (19) — их оценками (6), получим

$$\hat{p}(x_1, x_2, \ldots, x_{n-1}, x_n) = \frac{n(x_1)n(x_1, x_2)n(x_2, x_3)\ldots n(x_{n-1}, x_n)}{nn(x_1)n(x_2)\ldots n(x_{n-1})},$$

$$\hat{p}(\bar{x}_n, \bar{x}_{n-1}, \ldots, \bar{x}_2, \bar{x}_1) = \frac{n(\bar{x}_n)n(\bar{x}_n, \bar{x}_{n-1})n(\bar{x}_{n-1}, \bar{x}_{n-2})\ldots n(\bar{x}_2, \bar{x}_1)}{nn(\bar{x}_n)n(\bar{x}_{n-1})\ldots n(\bar{x}_2)},$$

поэтому из соотношений (1), (4) имеем (18).

Учитывая, что для модели цепей Маркова симметрия последовательностей оснований вытекает из симметрии пар, а из симметрии пар следует симметрия оснований, соотношения для пар $n(ij) = n(\bar{j}\bar{i})$, $i, j \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$, являются основными в записи генетической информации в ДНК.

С помощью модели цепей Маркова можно легко сгенерировать случайную последовательность, для которой будет выполняться симметрия вида (4), (10). На основе оценок переходных вероятностей (6), табл. 1 и программы псевдослучайных чисел строится случайная последовательность оснований, совпадающая по длине с хромосомой человека. Расчеты показали, что относительная разность между тройками оснований в (10) значительно меньше 1%. Таким образом, модель Маркова убедительно подтверждает симметрию коротких последовательностей оснований.

ЗАКЛЮЧЕНИЕ

В настоящей работе получены новые связывающие ограничения (8), (9) для пар AT, TA, CG и GC, которые не входили в соотношения (3) для пар оснований. Аналогичным образом получены новые ограничения (12)–(17) для троек оснований. Показано, что симметрия отдельных оснований является следствием симметрии пар оснований и соответственно симметрия пар оснований — следствием симметрии троек оснований. С помощью модели однородной цепи Маркова подтверждается, что симметрия последовательностей оснований вытекает из симметрии коротких последовательностей (пар оснований).

Решение сложных задач предсказания пространственной структуры белков показало, что если соотношения симметрии в записи генетической информации не выполняются, то байесовские процедуры распознавания на цепях Маркова не работают [3].

Полученные результаты открывают широкие возможности применения байесовских процедур на моделях цепей Маркова для распознавания свойств участков оснований (генов), расположенных на нитях ДНК.

СПИСОК ЛИТЕРАТУРЫ

1. Baisnée P.-F., Hampson S., Baldi P. Why are complementary DNA strands symmetric? // Bioinformatics. — 2002. — **18**, N 2. — P. 1021–1033.
2. Гупал А. М., Вагис А.А. Комплементарность оснований в хромосомах ДНК // Проблемы управления и информатики. — 2005. — № 5. — С. 90–94.
3. Гупал А.М., Сергиенко И.В. Оптимальные процедуры распознавания. — Киев: Наук. думка, 2008. — 232 с.
4. Anderson T.W., Goodman L.A. Statistical inference about Markov chains // Ann. Math. Stat. — 1957. — **28**. — P. 89–110.
5. Вейр Б. Анализ генетических данных. — М.: Мир, 1995. — 400 с.

Поступила 11.01.2011