УДК 519.217.2

А.А. ВАГИС, Н.А. ГУПАЛ

СИММЕТРИЯ В ЗАПИСИ ГЕНЕТИЧЕСКОЙ ИНФОРМАЦИИ В ДНК

Получены новые закономерности в записи пар, а также троек оснований по нитям в хромосомах ДНК. На основе модели однородной цепи Маркова показано, что комплементарность или симметрия высших порядков вытекают из комплементарности оснований и пар оснований.

Введение. Соотношения комплементарности или симметрии в записи оснований, подсчитанных по одной нити в хромосомах ДНК, исследовались в [1–3], в [1] содержится список цитируемой литературы по данному вопросу. В работах [2, 3] соотношения комплементарности приведены в виде коротких формул, что значительно облегчает и упрощает восприятие этих результатов и является основой построения математического аппарата с целью получения новых результатов. Статистический анализ подтвердил выполнение этих соотношений на геномах бактерий, растений, высших организмов (примерно сто геномов) в том числе и на ДНК человека. Соотношения комплементарности – яркий пример подтверждения законов симметрии в мире живой природы, однако до настоящего времени не выяснены причины, которые ответственны за появление этих законов в ДНК.

В работе получены новые закономерности в записи пар, а также троек оснований по нитям в хромосомах ДНК. На основе модели однородной цепи Маркова показано, что соотношения комплементарности или симметрии высших порядков для троек, четверок, а также коротких последовательностей оснований вытекают из комплементарности оснований и пар оснований.

ДНК имеет форму двойной спирали, информаця 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 на уровне отдельного гена.

Запись и считывание оснований у первой нити хромосомы ДНК выполняется слева направо в направлении $5' \to 3'$, а у комплементарной второй нити – в направлении $5' \to 3'$ справа налево (рисунок).

$5'$ _____AC_____GT___ $\to$ __$3'$ 1 - я нить

.....↕↕.....

$3'$ _____TG_____CA____ $\leftarrow$ __$5'$ 2 - я нить

РИСУНОК. Условная запись двух нитей хромосомы

Комплементарность в записи оснований по одной нити ДНК хромосомы означает, что выполняются приближенные соотношения

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

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 букв по двум нитям ДНК не следует, что количества букв A и T, а также C и G, подсчитанные по одной нити, совпадают между собой. Простой пример: на одной нити содержится 4 млн. букв A, 3 млн. букв C, 2 млн. букв G и 1 млн. букв T, тогда на второй нити находится соответственно 4 млн. букв T, 3 млн. букв G, 2 млн. букв C и 1 млн. букв A. Таким образом, комплементарность по двум нитям выполняется, а по одной нити нет.

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

$$n(\mathrm{A},1) = n(\mathrm{A},2), \quad n(\mathrm{T},1) = n(\mathrm{T},2),$$

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

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

Расчеты показали, что для пар оснований выполняются следующие соотношения комплементарности:

$$n(\mathrm{AC}) = n(\mathrm{GT}), \quad n(\mathrm{AG}) = n(\mathrm{CT}),$$

$$n(\mathrm{TC}) = n(\mathrm{GA}), \quad n(\mathrm{TG}) = n(\mathrm{CA}),$$

$$n(\mathrm{AA}) = n(\mathrm{TT}), \quad n(\mathrm{CC}) = n(\mathrm{GG}), \tag{3}$$

или короче в виде формулы

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

где $i, j \in \{\mathrm{A},\mathrm{C},\mathrm{G},\mathrm{T}\}$, $\bar{\mathrm{A}} = \mathrm{T}$, $\bar{\mathrm{C}} = \mathrm{G}$, $\bar{\mathrm{T}} = \mathrm{A}$, $\bar{\mathrm{G}} = \mathrm{C}$. Заметим, что пары AT, TA, CG и GC не присутствуют в (3), поскольку они приводят к тавтологии (в табл. 1 приведены количества пар оснований).

Симметрия пар оснований. 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.

ТАБЛИЦА 1. Геном человека

Пары букв	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 симметрии (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 (см. рисунок). 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

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

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

где $i \in \{\mathrm{A},\mathrm{C},\mathrm{G},\mathrm{T}\}$. Для основания A из (7) получаем новое соотношение для пар AT, TA, которые не входят в (3):

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

Для основания C из (7) – новое соотношение для пар CG и GC

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

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

Кодоны (тройки оснований) связаны следующими соотношениями комплементарности:

$$n(i, j, k) \approx n(\bar{k}, \bar{j}, \bar{i}), \qquad (10)$$

где $n(i, j, k)$ – число троек оснований (i, j, k), $(\bar{k}, \bar{j}, \bar{i})$ – антикодон кодона (i, j, k). Для 64 триплетов получаем 32 соотношения (10) типа кодон – антикодон (табл. 2).

Из соотношений (10) вытекает, что для однородной цепи Маркова оценки вероятностей троек оснований (i, j, k) и $(\bar{k}, \bar{j}, \bar{i})$ совпадают:

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

где n – длина хромосомы. Модель однородной цепи Маркова интересна тем, что для нее комплементарность троек оснований (10) вытекает из комплементарности оснований (1) и комплементарности пар оснований (4).

ТАБЛИЦА 2. Количество кодонов в хромосоме 6 генома человека

Кодон	Число	Кодон	Число	Кодон	Число	Кодон	Число
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

Симметрия троек оснований. 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), $i, j \in \{\mathrm{A},\mathrm{C},\mathrm{G},\mathrm{T}\}$, как и для (7), справедливы соотношения

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

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

Для шести пар (3), используя соотношения (10), выводим формулы

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

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

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

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

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

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

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. Заметим, что соотношения (12) – (17) выполняются для модели однородной цепи Маркова. Для подтверждения (12) покажем, что

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

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

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

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

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

Остается воспользоваться уже выведенной формулой (8). Для обоснования (13), учитывая формулу (7):

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

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

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

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

Соотношения комплементарности и симметрии для коротких последовательностей оснований также подтверждаются для модели однородной цепи Маркова. Этот результат вытекает из следующего важного утверждения.

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

$$\hat{p}(x_1, x_2, ..., x_{n-1}, x_n) = \hat{p}(\overline{x_n}, \overline{x_{n-1}}, ..., \overline{x_2}, \overline{x_1}). \quad (18)$$

Заключение. В работе выведены новые соотношения (8), (9) в записи пар AT, TA, CG и GC, которые отсутствовали в ранее полученных соотношениях комплементарности (3) для пар оснований. Аналогично получены новые формулы зависимостей (12)–(17) для троек оснований. С помощью модели однородной цепи Маркова удалось показать, что комплементарность или симметрия высших порядков для коротких последовательностей оснований вытекают из комплементарности для одиночных оснований и пар оснований. Полученные результаты открывают широкие возможности применения байесовских процедур на моделях цепей Маркова для распознавания последовательностей оснований, а также белок-кодирующих участков, расположенных на нитях ДНК.

О.А. Вагіс, М.А. Гупал

СИМЕТРІЯ У ЗАПИСУ ГЕНЕТИЧНОЇ ІНФОРМАЦІЇ В ДНК

Отримано нові формули залежностей для пар і трійок основ в ДНК. За допомогою апарату ланцюгів Маркова показано, що комплементарність і симетрія вищих порядків для послідовностей основ витікає із комплементарності і симетрії для пар основ.

A.A. Vagis, N.A. Gupal

SYMMETRY IN RECORDING OF GENETIC INFORMATION IN DNA

New formulas for dependences are obtained for the pairs and triples of the bases in DNA. Using Markov chains, it is shown that complementarity and symmetry of higher orders for the sequences of bases follows from complementarity and symmetry for the pair of bases.

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. 153–157.
3. *Гупал А.М., Сергиенко И.В.* Оптимальные процедуры распознавания. – Киев: Наук. думка, 2008. – 232 с.

Получено 16.12.2010

Об авторах:

Вагис Александра Анатольевна,
кандидат физико-математических наук, старший научный сотрудник
Института кибернетики имени В.М. Глушкова НАН Украины,

Гупал Никита Анатольевич,
младший научный сотрудник Института кибернетики имени В.М. Глушкова НАН Украины.