*Дискуссии*

УДК 519.217.2

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

Институт кибернетики им. В.М. Глушкова НАН Украины, Киев
E-mail: gupal_anatol@yahoo.com

# СООТНОШЕНИЯ КОМПЛЕМЕНТАРНОСТИ В ЗАПИСИ ОСНОВАНИЙ ПО ОДНОЙ НИТИ ДНК

*Установлены новые комплементарные закономерности по записи генетической информации в геноме человека и других исследованных геномов. Частоты комплементарных букв (азотистых оснований) в одной цепи ДНК совпадают между собой для всех хромосом исследованных геномов, т.е. закономерность, аналогичная правилу Чаргаффа (и принципу комплементарности Уотсона-Крика), выполняется не только для двухцепочечной ДНК, но и для каждой отдельной нити. Полученные результаты дополнят современные представления относительно записи генетической информации в ДНК, а также прохождения таких важных процессов, как репликация ДНК и образование компактной формы укладки хромосом в ядрах клеток организма и т.д.*



**Введение.** Несмотря на то, что ДНК относительно проста и хорошо изучена химически, структура генома человека чрезвычайно сложна и не все его функции известны. На текущий момент длина законченной геномной последовательности составляет 2851 млн нуклеотидов и содержит 341 пробел общим размером 225 млн оснований. Геном человека включает приблизительно 20–30 тысяч белок-кодирующих генов. В работе [1] приведены сведения о законченных последовательностях и размерах пробелов для каждой хромосомы в геноме человека. Числовые расчеты проводились на последовательностях хромосом, характеристики которых соответствуют данным, указанным в [1].

ДНК имеет форму двойной спирали, информация записана в четырехбуквенном алфавите оснований $A, C, G, T$; $C–G, A–T$ – комплементарные пары оснований, связывающие две цепи.

Хромосомы – функционально интегрированные участки ДНК, в них содержится информация относительно тысяч генов, поэтому расчеты проводятся на уровне всей хромосомы, а не на уровне отдельного гена.

**Комплементарность оснований.** Заметим, что из комплементарности пар букв по двум нитям ДНК не следует, что количество букв $A$ и $T$, а также $C$ и $G$, подсчитанные по одной нити, совпадает между собой. Простой пример: на одной нити содержится 3 млн букв $A$ и $C$ и по 1 млн букв $T$ и $G$, тогда на второй нити находится соответственно по 3 млн букв $T$ и $G$ и по 1 млн букв $A$ и $C$. Таким образом, комплементарность по двум нитям выполняется, а по одной нити нет.

Хромосомы имеют разную длину, поэтому анализируются частоты, а не отдельные значения оснований (пар оснований и отдельных $n$-ок). Частота буквы $j$, $j \in \{A,C,G,T\}$, есть

$$\frac{m(j)}{m},$$

где $m(j)$ – число букв $j$, $m$ – длина хромосомы. Вычисления показали, что частоты коплементарных оснований $A$ и $T$, а также $C$ и $G$, подсчитанные по одной нити ДНК, совпадают на всех хромосомах (геном человека, шимпанзе, рыба Tetraodon [2] (табл. 1). Поэтому по свойству комплементарности оснований для каждой из двух нитей хромосом выполняются со-

Таблица 1

**Частоты оснований *A*, *C*, *G*, *T*, подсчитанные на хромосомах трех геномов**

| Хромосома | Геном | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | человека | | | | шимпанзе | | | | Tetraodon | | | |
| | A | C | G | T | A | C | G | T | A | C | G | T |
| 1 | 0,291 | 0,209 | 0,209 | 0,292 | 0,292 | 0,207 | 0,207 | 0,293 | 0,278 | 0,223 | 0,223 | 0,276 |
| 2 | 0,299 | 0,201 | 0,201 | 0,299 | 0,302 | 0,198 | 0,198 | 0,303 | 0,272 | 0,228 | 0,228 | 0,272 |
| 3 | 0,301 | 0,198 | 0,198 | 0,302 | 0,310 | 0,190 | 0,190 | 0,310 | 0,267 | 0,234 | 0,233 | 0,266 |
| 4 | 0,309 | 0,191 | 0,191 | 0,309 | 0,303 | 0,197 | 0,197 | 0,303 | 0,269 | 0,231 | 0,232 | 0,268 |
| 5 | 0,302 | 0,197 | 0,198 | 0,303 | 0,303 | 0,197 | 0,197 | 0,303 | 0,277 | 0,223 | 0,223 | 0,278 |
| 6 | 0,302 | 0,198 | 0,198 | 0,302 | 0,299 | 0,201 | 0,201 | 0,299 | 0,263 | 0,235 | 0,235 | 0,267 |
| 7 | 0,296 | 0,204 | 0,204 | 0,297 | 0,300 | 0,200 | 0,200 | 0,300 | 0,273 | 0,228 | 0,228 | 0,272 |
| 8 | 0,299 | 0,201 | 0,201 | 0,299 | 0,292 | 0,207 | 0,207 | 0,293 | 0,272 | 0,228 | 0,228 | 0,272 |
| 9 | 0,293 | 0,207 | 0,207 | 0,293 | 0,293 | 0,207 | 0,207 | 0,293 | 0,269 | 0,231 | 0,232 | 0,268 |
| 10 | 0,292 | 0,208 | 0,208 | 0,292 | 0,294 | 0,205 | 0,205 | 0,295 | 0,266 | 0,233 | 0,233 | 0,267 |
| 11 | 0,292 | 0,208 | 0,208 | 0,292 | 0,295 | 0,205 | 0,205 | 0,295 | 0,269 | 0,229 | 0,230 | 0,272 |
| 12 | 0,296 | 0,204 | 0,204 | 0,296 | 0,296 | 0,204 | 0,204 | 0,296 | 0,269 | 0,229 | 0,230 | 0,272 |
| 13 | 0,307 | 0,193 | 0,193 | 0,308 | 0,302 | 0,197 | 0,197 | 0,303 | 0,271 | 0,229 | 0,229 | 0,271 |
| 14 | 0,294 | 0,204 | 0,205 | 0,297 | 0,306 | 0,193 | 0,193 | 0,308 | 0,268 | 0,232 | 0,232 | 0,268 |
| 15 | 0,289 | 0,211 | 0,211 | 0,289 | 0,295 | 0,203 | 0,204 | 0,298 | 0,263 | 0,237 | 0,237 | 0,263 |
| 16 | 0,275 | 0,223 | 0,224 | 0,277 | 0,291 | 0,209 | 0,208 | 0,291 | 0,271 | 0,230 | 0,230 | 0,270 |
| 17 | 0,272 | 0,228 | 0,227 | 0,273 | 0,302 | 0,198 | 0,198 | 0,302 | 0,275 | 0,224 | 0,224 | 0,277 |
| 18 | 0,301 | 0,199 | 0,199 | 0,301 | 0,276 | 0,223 | 0,223 | 0,278 | 0,271 | 0,230 | 0,230 | 0,269 |
| 19 | 0,258 | 0,242 | 0,242 | 0,259 | 0,269 | 0,232 | 0,231 | 0,268 | 0,261 | 0,239 | 0,239 | 0,260 |
| 20 | 0,278 | 0,220 | 0,221 | 0,280 | 0,259 | 0,240 | 0,240 | 0,260 | 0,277 | 0,223 | 0,221 | 0,279 |
| 21 | 0,297 | 0,204 | 0,205 | 0,294 | 0,279 | 0,219 | 0,220 | 0,282 | 0,269 | 0,232 | 0,232 | 0,267 |
| 22 | 0,261 | 0,240 | 0,240 | 0,260 | 0,297 | 0,204 | 0,204 | 0,295 | | | | |
| 23 | | | | | 0,262 | 0,239 | 0,238 | 0,261 | | | | |
| X | 0,302 | 0,197 | 0,197 | 0,303 | 0,303 | 0,196 | 0,196 | 0,304 | | | | |
| Y | 0,299 | 0,199 | 0,200 | 0,301 | 0,300 | 0,198 | 0,198 | 0,302 | | | | |

отношения

$$\frac{m(A)}{m} = \frac{m(T)}{m}, \qquad \frac{m(C)}{m} = \frac{m(G)}{m}. \qquad (1)$$

Из (1) вытекает важный вывод о том, что молекулярная масса каждой нити ДНК одинакова. Если бы частота основания *A* превосходила частоту основания *T* (аналогично относительно оснований *C* и *G*), то в силу огромной длины ДНК масса одной нити превосходила бы массу другой нити, т.е. молекула ДНК была бы сильно перегружена и не была бы устойчивой.

Grimwood J. [3] отмечает, что хромосома 19 в геноме человека имеет самую высокую плотность генов среди всех хромосом, что соответствует высокому содержанию оснований $C + G$ в хромосоме. Из табл. 1 видно, что таким свойством обладают хромосомы 16, 17, 19, 20, 22 в геноме человека, а также хромосомы 18, 19, 20, 21, 23 в геноме шимпанзе. Анализ других геномов (мышь, крыса, цыпленок, *C. elegans*, *Arabidopsis* и др.) показал, что в этих геномах также выполняется свойство комплементарности оснований (1) (в геноме цыпленка для хромосом длиной свыше 900 тысяч оснований). Заметим, что диапазон изменения по хромосомам частоты отдельного основания составляет величину 0,051 для генома человека, 0,051 – для шимпанзе и 0,017 – для генома Tetraodon.

**Комплементарность пар оснований.** Частоты пар оснований вычисляются по формуле

$$\frac{m(ij)}{m(i)}, \qquad (2)$$

где $m(ij)$ – число пар $(ij)$, $i, j \in \{A,C,G,T\}$, $m(i)$ – число букв $i$ в цепи хромосомы. Соотношения (2) являются оценками переходных вероятностей для стационарных цепей Маркова [4]. С помощью решения задач распознавания гипотез

показано, что однородная цепь Маркова наилучшим образом (по сравнению с цепями более высоких порядков) соответствует данным, записанным в хромосомах [5, 6].

В табл. 2 приведены частоты пар оснований для трех геномов. Диапазон изменения частот пар оснований по хромосомам составляет для генома человека 0,057 (он максимален для пары *CA*, которая имеет максимальную частоту), для генома шимпанзе – 0,055, для генома Tetraodon – 0,033. Минимальной частотой обладает пара оснований *CG*.

Замечательная особенность поведения частот пар оснований состоит в том, что для всех хромосом этих геномов выполняются соотношения

$$\frac{m(AA)}{m(A)} = \frac{m(TT)}{m(T)}, \qquad \frac{m(CC)}{m(C)} = \frac{m(GG)}{m(G)}. \quad (3)$$

Расчеты показали, что они выполняются и для других указанных геномов.

Интересная особенность поведения частот пар оснований заключается в том, что вторая комплементарная нить в направлении 5′–3′ (это направление противоположно направлению 5′–3′ первой нити) имеет такие же частоты (2), что и исходная первая нить. В силу комплементарности пар оснований *A* и *T* (*C* и *G*) и соотношений (1) этот результат очевиден для пар *AA, AT, CC, CG, GC, GG, TA* и *TT*. Совпадение частот для остальных пар оснований *AC, AG, CA, CT, GA, GT, TC* и *TG* легко проверяется расчетами на компьютере. Отсюда следует, что вероятности двух противоположных нитей хромосом, подсчитанные в модели однородной цепи Маркова, совпадают.

**Комплементарность последовательностей одинаковых оснований.** В геномах обнаружены другие интересные регулярности относительно повторов одинаковых комплементарных букв. Компьютер подсчитывал число изолированных последовательностей, состоящих из одинаковых букв *A, T, C, G*. Изолированная буква *A* не входит в состав пар *AA*, троек *AAA* и т.д., пара *AA* не входит в состав троек *AAA*, четверок *AAAA* и т.д. Таким образом, последовательности, состоящие из одинаковых букв *A*, не пересекаются и в сумме дают общее число букв *A* в хромосоме. То же самое относится к последовательностям, состоящим из одинаковых букв *T, C, G*. В табл. 3 приведены данные о количе-

Таблица 2

**Частоты пар оснований**

| Пары букв | Геном | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | человека | | | | | | шимпанзе | | | | | | Tetraodon | | | | | |
| | Хромосома | | | | | | | | | | | | | | | | | |
| | 1 | 9 | 16 | 22 | X | Y | 1 | 9 | 16 | 23 | X | Y | 1 | 5 | 9 | 13 | 20 | 21 |
| AA | 0,33 | 0,33 | 0,31 | 0,29 | 0,34 | 0,33 | 0,33 | 0,33 | 0,33 | 0,29 | 0,34 | 0,34 | 0,32 | 0,32 | 0,31 | 0,32 | 0,33 | 0,32 |
| AC | 0,17 | 0,17 | 0,18 | 0,20 | 0,17 | 0,17 | 0,17 | 0,17 | 0,17 | 0,20 | 0,17 | 0,17 | 0,21 | 0,21 | 0,22 | 0,21 | 0,20 | 0,21 |
| AG | 0,24 | 0,24 | 0,26 | 0,29 | 0,23 | 0,23 | 0,24 | 0,24 | 0,24 | 0,29 | 0,22 | 0,23 | 0,24 | 0,23 | 0,25 | 0,25 | 0,24 | 0,25 |
| AT | 0,26 | 0,26 | 0,24 | 0,22 | 0,27 | 0,27 | 0,26 | 0,26 | 0,25 | 0,22 | 0,27 | 0,27 | 0,23 | 0,22 | 0,22 | 0,22 | 0,23 | 0,22 |
| CA | 0,35 | 0,35 | 0,34 | 0,32 | 0,37 | 0,37 | 0,35 | 0,35 | 0,35 | 0,32 | 0,37 | 0,37 | 0,35 | 0,35 | 0,33 | 0,33 | 0,34 | 0,33 |
| CC | 0,26 | 0,26 | 0,28 | 0,30 | 0,25 | 0,25 | 0,26 | 0,26 | 0,26 | 0,29 | 0,24 | 0,24 | 0,23 | 0,23 | 0,24 | 0,24 | 0,23 | 0,24 |
| CG | 0,05 | 0,05 | 0,06 | 0,07 | 0,04 | 0,04 | 0,05 | 0,05 | 0,05 | 0,07 | 0,04 | 0,04 | 0,12 | 0,12 | 0,14 | 0,14 | 0,13 | 0,15 |
| CT | 0,34 | 0,34 | 0,32 | 0,31 | 0,35 | 0,34 | 0,34 | 0,34 | 0,34 | 0,31 | 0,35 | 0,35 | 0,30 | 0,30 | 0,29 | 0,29 | 0,30 | 0,28 |
| GA | 0,29 | 0,29 | 0,27 | 0,25 | 0,30 | 0,30 | 0,29 | 0,29 | 0,29 | 0,35 | 0,30 | 0,30 | 0,27 | 0,27 | 0,27 | 0,27 | 0,27 | 0,26 |
| GC | 0,21 | 0,21 | 0,22 | 0,24 | 0,20 | 0,20 | 0,21 | 0,21 | 0,21 | 0,24 | 0,20 | 0,20 | 0,24 | 0,24 | 0,24 | 0,24 | 0,25 | 0,25 |
| GG | 0,26 | 0,26 | 0,28 | 0,30 | 0,25 | 0,25 | 0,26 | 0,26 | 0,26 | 0,29 | 0,24 | 0,24 | 0,23 | 0,23 | 0,24 | 0,24 | 0,23 | 0,24 |
| GT | 0,24 | 0,24 | 0,23 | 0,21 | 0,26 | 0,26 | 0,24 | 0,24 | 0,24 | 0,21 | 0,26 | 0,26 | 0,26 | 0,26 | 0,25 | 0,25 | 0,25 | 0,25 |
| TA | 0,22 | 0,22 | 0,20 | 0,18 | 0,23 | 0,22 | 0,22 | 0,22 | 0,22 | 0,18 | 0,23 | 0,22 | 0,18 | 0,18 | 0,17 | 0,17 | 0,18 | 0,17 |
| TC | 0,20 | 0,20 | 0,22 | 0,23 | 0,19 | 0,20 | 0,20 | 0,20 | 0,20 | 0,23 | 0,19 | 0,19 | 0,22 | 0,22 | 0,23 | 0,23 | 0,22 | 0,23 |
| TG | 0,25 | 0,25 | 0,27 | 0,30 | 0,24 | 0,24 | 0,25 | 0,25 | 0,25 | 0,29 | 0,24 | 0,25 | 0,28 | 0,28 | 0,29 | 0,28 | 0,27 | 0,28 |
| TT | 0,33 | 0,33 | 0,31 | 0,29 | 0,34 | 0,33 | 0,33 | 0,33 | 0,33 | 0,29 | 0,34 | 0,34 | 0,32 | 0,32 | 0,31 | 0,32 | 0,33 | 0,32 |

Таблица 3

**Количество последовательностей, состоящих из одинаковых букв, в хромосоме 2 генома человека**

| Размер *n*-ки | A | T | C | G |
|---|---|---|---|---|
| 1 | 32 885 475 | 32 885 555 | 26 877 090 | 26 900 089 |
| 2 | 8 802 666 | 8 823 505 | 6 695 063 | 6 700 669 |
| 3 | 3 452 571 | 3 465 217 | 1 730 704 | 1 730 727 |
| 4 | 1 330 971 | 1 335 874 | 416 239 | 417 181 |
| 5 | 502 189 | 505 290 | 96 319 | 96 463 |
| 10 | 8179 | 8255 | 131 | 131 |
| 15 | 2525 | 2604 | 8 | 10 |
| 18 | 1389 | 1434 | 3 | 1 |
| 20 | 1044 | 1022 | 1 | 1 |
| 24 | 635 | 608 | 1 | 1 |
| 40 | 18 | 15 | 0 | 0 |
| 50 | 3 | 3 | 0 | 0 |
| 62 | 1 | 1 | 0 | 0 |

Таблица 4

**Количество последовательностей, состоящих из одинаковых букв, в хромосоме 7 генома Tetraodon**

| Размер *n*-ки | A | T | C | G |
|---|---|---|---|---|
| 1 | 1 295 197 | 1 292 165 | 1 294 911 | 1 292 375 |
| 2 | 303 520 | 303 136 | 311 692 | 311 146 |
| 3 | 132 007 | 130 530 | 63 221 | 64 284 |
| 4 | 45 534 | 45 462 | 15 496 | 15 632 |
| 5 | 16 085 | 15 887 | 4375 | 4570 |
| 10 | 479 | 420 | 63 | 52 |
| 15 | 89 | 78 | 12 | 12 |
| 20 | 17 | 16 | 3 | 2 |
| 25 | 2 | 3 | 1 | 1 |
| 30 | 3 | 3 | 0 | 1 |
| 35 | 1 | 0 | 0 | 0 |

ствах последовательностей, состоящих из одинаковых букв *A, T,C, G*, в хромосоме 2 генома человека.

Отсюда можно сделать вывод о том, что выполняются соотношения

$$n(A...A) \sim n(T...T),$$
$$n(C...C) \sim n(G...G). \quad (4)$$

Соотношения (4) были подтверждены для остальных хромосом исследуемых геномов. Мы не ставим знак равенства в соотношениях (4) из-за наличия пропусков в геномах и точности секвенирования геномов. Заметим, что число различных вариантов последовательностей, состоящих из 20 букв, составляет $4^{20} = 2^{40} \sim 10^{12}$, а для 50 букв — $4^{50} = 2^{100} \sim 10^{30}$. Очевидно, что вероятность того, что мы случайно обнаружили справедливость выполнения соотношений (4) для всех *n*-ок последовательностей, составляет бесконечно малую величину.

**Выводы**. Установлены новые соотношения комплементарности относительно записи генетической информации в геноме человека и других исследуемых геномах. Показано, что частоты комплементарных оснований, подсчитанные по одной нити ДНК, совпадают между собой для всех хромосом исследуемых геномов. Расчеты показали, что количество *n*-ок одинаковых комплементарных оснований также совпадает на каждой отдельной нити ДНК хромосом. Полученные результаты дополнят современные представления относительно записи генетической информации в ДНК, а также прохождения таких важных процессов, как репликация ДНК и образования компактной формы укладки хромосом.

*РЕЗЮМЕ.* Встановлено нові комплементарні закономірності щодо запису генетичної інформації в геномі людини та інших досліджуваних геномів. Частоти комплементарних літер (азотистих основ) в одному ланцюзі ДНК співпадають між собою для всіх хромосом досліджуваних геномів, тобто закономірність, аналогічна правилу Чаргаффа (і принципу комплементарності Уотсона-Кріка), виконується не тільки для дволанцюгової ДНК, але і для кожної окремої низки. Отримані результати доповнять сучасні уявлення відносно запису генетичної інформації в ДНК, а також про перебіг таких важливих процесів, як реплікація ДНК та утворення компактної форми укладки хромосом ДНК в ядрах клітин організму тощо.

*SUMMARY.* We have determined the new complementary principles in encoding bases on DNA chain in chromosomes of human genome and some other investigated genomes. The obtained results show that regularity analogues to Chargaff rule (or complementary principle of Watson-Crick) holds not only for two-chain DNA spiral but even for each separate chain. Moreover, we revealed other remarkable regularities concerning repeating sequences of sets of identical letters (bases). On the basis of obtained statistical data one can draw a conclusion that there exist some strict rules of forming DNA structure valid for all species.

The obtained results will significantly improve our present view of encoding genetic information and also data on DNA replication process as well as formation of compact shape of chromosome packing.

## СПИСОК ЛИТЕРАТУРЫ


1. *The International* Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome // Nature. – 2004. – **431**. – P. 931–945.
2. *Jaillon O. et al.* Genome duplication in the fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype // Nature. – 2004. – **431**. – P. 946–957.
3. *Grimwood J. et al.* The DNA sequence and biology of human chromosome 19 // Nature. – 2004. – **428**. – P. 529–535.
4. *Anderson T.W., Goodman L.A.* Statistical inference about Markov chains // Ann. Math. Statist. – 1957. – **28**. – P. 89–110.
5. *Сергиенко И.В., Гупал А.М., Воробьев А.С., Вагис А.А.* Математическая модель генома // Проблемы управления и информатики. – 2004. – № 2. – С. 124–129.
6. *Сергиенко И.В., Гупал А.М.* Статистический анализ генома // Цитология и генетика. – 2004. – **38**, № 4. – С. 76–81.


Поступила 01.07.05