КОМП'ЮТЕРНІ ЗАСОБИ, МЕРЕЖІ ТА СИСТЕМИ

S.V. Vasylyev, A.M. Gupal

FEATURES OF BASES DISPOSITION IN DNA RESEARCH USING CLUSTER COMPUTER

It is determined the new complementary principles in encoding bases by one chain in DNA chromosomes of human genome and other investigated genomes. On the basis of obtained statistical data one can draw a conclusion that there exist strict rules of forming DNA structure valid for all species.
The obtained results will significantly improve modern present view of encoding genetic information.

Установлены новые соотношения комплементарности относительно записи оснований по одной нити хромосом ДНК. Полученные результаты существенно дополняют современные представления о записи генетической информации в ДНК.

УДК 519.217.2

С.В. ВАСИЛЬЕВ, А.М. ГУПАЛ

ИССЛЕДОВАНИЕ ОСОБЕННОСТЕЙ ЗАПИСИ ОСНОВАНИЙ В ДНК НА КЛАСТЕРНОМ КОМПЬЮТЕРЕ

Введение. В работах [1, 2] выведены соотношения комплементарности в записи оснований, подсчитанные по одной нити ДНК. Эти соотношения были приведены для частот оснований, а также частот *n*-к оснований. При переходе от частот оснований к вычислению оснований, пар оснований, отдельных n-ок оснований обнаружен ряд новых замечательных соотношений комплементанности.

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 – непрерывные участки ДНК, в которых содержится информация относительно тысяч генов. Поэтому расчеты проводятся на уровне всей хромосомы, а не на уровне отдельного гена. Проведен статистический анализ более 40 геномов, из них около 30 геномов бактерий.

Комплементарность оснований. 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. Рассмотрим 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.

Хромосомы имеют разную длину, поэтому анализируются частоты, а не отдельные значения оснований (пар оснований и отдельных *n*-к).

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 количества оснований A и T, а также C и G, подсчитанные по одной нити ДНК, практически совпадают на всех хромосомах [1, 2]. Поэтому по свойству комплементарности оснований для каждой из двух нитей хромосом выполняются соотношения

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

где $n(j)$ – число букв j, $j \in \{A,C,G,T\}$.

Мы не ставим знак равенства в (1) по следующим причинам: геном – динамическая структура, постоянно подверженная модификациям в процессе эволюции; геномы всех живущих на Земле людей отличаются друг от друга; наличие пробелов в геноме человека и в большинстве других геномов; имеет место определенная точность секвенирования геномов [3].

Комплементарность пар оснований. Частоты пар оснований вычисляются по формулам

$$\hat{p}(ij) = \frac{m(ij)}{m(i)}, \tag{2}$$

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 работе [1] показано, что для всех хромосом исследуемых геномов выполняются соотношения

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

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. Отсюда следует, что вероятности двух противоположных нитей хромосом, подсчитанные в модели однородной цепи Маркова, совпадают.

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

Если предположить, что запись информации во второй нити проводится по такой же схеме, что и у первой нити, то количество пар AC в первой нити

должно совпадать с количеством пар *GT*, поскольку в силу комплементарности паре *AC* на второй нити соответствует пара *GT* на первой нити. То же самое касается остальных пар.

$$5' \_\_\_\_\_ AC \_\_\_\_\_ GT \_\_\_\_\_ \rightarrow \_\_ 3'$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \updownarrow\updownarrow \ldots\ldots\ldots\ldots\ldots\ldots$$
$$3' \_\_\_\_\_ CA \_\_\_\_\_ \leftarrow \_\_ 5'$$

РИСУНОК

Расчеты подтвердили это предположение. Для пар букв выполняются следующие соотношения комплементарности:

$$n(AC) \sim n(GT), \quad n(AG) \sim n(CT),$$
$$n(TC) \sim n(GA), \quad n(TG) \sim n(CA),$$
$$n(AA) \sim n(TT), \quad n(CC) \sim n(GG). \qquad (4)$$

Заметим, что пары *AT, TA, CG,* и *GC* не присутствуют в (4), поскольку они сами себе антикомплементарны (табл. 1).

ТАБЛИЦА 1

Геном человека. Хромосомы

Пары букв	1	3	6	10	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

При переходе от частот к количествам пар получаем вместо двух шесть соотношений комплементарности. Для частот пар их получить было нельзя из-за того, что в частотах (2) присутствуют разные основания.

Кодоны (тройки оснований) связаны следующими соотношениями комплементарности:

$$n(ij...k) = n(\bar{k}...\bar{j}\,\bar{i}), \tag{5}$$

где $i, j, k \in \{A, C, G, T\}$, $\bar{A} = T$, $\bar{C} = G$, $\bar{T} = A$, $\bar{G} = C$, $(\bar{k}...\bar{j}\,\bar{i})$ – антикодон кодона ($ij...k$) (табл. 2).

ТАБЛИЦА 2

Количество кодонов в хромосоме шесть 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	3 412 539	*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	*GGT*	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

Термином «*n*-ка» (основания) мы будем называть число, обозначающее, сколько раз встречается данная подпоследовательность в общей последовательности (например, *n*(*AAA*), *n*(*ACGT*), *n*(*TT*) и т.д.).

Для нечетных последовательностей каждая *n*-ка имеет антикомплементарную *n*-ку, исключений в этом случае нет. Для 64 триплетов получаем 32 соотношения: кодон – антикодон.

Каждая четверка оснований имеет антикомплементраную ей четверку, кроме 16 исключений, которые получаются из пар *AT, TA, CG,* и *GC* вставками в середину каждой из этих пар. Вычисления показали, что аналогичные соотноше-

ния комплементарности выполняются для возрастающих n-к (расчеты не приводятся из-за больших размеров таблиц).

Расчет таблиц проводился с помощью мультипроцессорной системы «СКИТ», разработанной в Институте кибернетики им. В.М. Глушкова НАН Украины. Она представляет собой 32-процессорный 16-узловой кластер на основе микропроцессоров Intel Xeon 2,67 ГГц. Особенностью поставленной задачи является большой объем входных данных (около 3,5 Гб). Входные данные были предварительно размещены на файловом сервере кластера, доступ к которому организован на основе высокоскоростных соединений сети SCI. Программа написана на языке C++ и откомпилирована компиляторами gcc (Linux) и Borland C (Windows). Для параллельного программирования использована система обмена сообщениями MPI.

Задача была разбита на 24 подзадачи (по числу хромосом), которые выполнялись в параллельном режиме без обмена данных между процессами, что обеспечивает минимальную затрату времени на обмен сообщениями между процессорами и, соответственно, максимальное быстродействие системы. Выходными данными являются количество всех пар (троек, четверок, и т.д.) нуклеотидов, а также их частоты.

Время подсчета данных для всех 24 хромосом составила 250 с (время подсчета 1-й – самой большой из хромосом). Для сравнения – время счета на настольном ПК составила 1990 с. Конфигурация: процессор Athlon 950 МГц/256 Мб ОЗУ/HDD 40Gb 5600. 24 виртуальных процессора эмулированы на одном физическом при использовании библиотеки mpich под Windows XP. Время подсчета одной 1-й хромосомы на такой системе составило 380 с.

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 разной длины, состоящие только из букв A, не пересекаются и в сумме дают общее число букв A в хромосоме. То же самое относится к последовательностям, состоящим из букв T, C, G. В табл. 3 приведены данные о количествах подпоследовательностей (изолированных n-х), состоящих из букв A, T, C, G, в хромосоме два генома человека.

Отсюда можно сделать вывод о том, что выполняются следующие соотношения:

$$n(A...A) = n(T...T), \;\; n(C...C) = n(G...G). \qquad (6)$$

Соотношения (6) были подтверждены для остальных хромосом исследуемых 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 (6) для всех *n*-к последовательностей, составляет бесконечно малую величину.

ТАБЛИЦА 3

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	8 179	8 255	131	131
15	2 525	2 604	8	10
18	1 389	1 434	3	1
20	1 044	1 022	1	1
24	635	608	1	1
40	18	15	0	0
50	3	3	0	0
62	1	1	0	0

Выводы. Установлены новые соотношения комплементарности в хромосомах ДНК, которые существенно дополняют современные представления относительно записи генетической информации в ДНК. На основе полученных статистических данных можно сделать вывод о том, что существуют строгие правила формирования структуры ДНК, которые справедливы для всех видов организмов.

1. *Сергиенко И.В., Гупал А.М., Вагис А.А.* Соотношения комплементарности в записи оснований по одной нити в хромосомах ДНК // Проблемы управления и информатики. – 2005. – № 4. – С. 153 –157.
2. *Гупал А.М., Вагис А.А.* Комплементарность оснований в хромосомах ДНК // Проблемы управления и информатики. – 2005. – № 5. – С. 90 –94.
3. The International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome // Nature. – 2004. – **431**. – P. 931–945.
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. – № 6. – С. 68 –74.
6. *Сергиенко И.В., Гупал А.М., Вагис А.А.* Соотношения комплементарности в записи оснований по одной нити в ДНК // Цитология и генетика. – 2005. – № 6. – С. 83 –87.

Получено 20.02.2006