СИММЕТРИЯ В ЗАПИСИ ГЕНЕТИЧЕСКОЙ ИНФОРМАЦИИ В ДНК

Анатолий Гупал, Александра Вагис

***Аннотация**: Показано, что для нитей ДНК возможно два вида симметрии, но в природе реализован один более эффективный способ записи и считывания информации. Доказано, что из симметрии последовательностей оснований вытекает симметрия коротких последовательностей, в том числе отдельных оснований. На основе модели цепей Маркова показано, что симметрия последовательностей оснований вытекает из симметрии пар оснований.*

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

ACM Classification Keywords: G.3 Probability and statistics.

Введение

Симметрия в записи оснований, подсчитанных по нитям в хромосомах ДНК, исследовалась в работах [1, 2]. Соотношения симметрии приведены в виде коротких формул, что значительно упрощает восприятие этих результатов и является основой построения математического аппарата для получения новых результатов. Статистический анализ подтвердил выполнение соотношений симметрии на геномах бактерий, растений, высших организмов (примерно сто геномов), в том числе и на ДНК человека [1, 2]. 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 [3] 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. Выведены новые связывающие ограничения для пар и троек оснований. На основе модели цепей Маркова показано, что симметрия для троек и коротких последовательностей оснований вытекает из симметрии пар оснований. В настоящей работе исследованы свойства двух видов симметрии для противоположной и одинаковой полярности цепей ДНК.

1. Противоположная полярность цепей ДНК

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 на уровне отдельного гена. Запись и считывание оснований у первой нити хромосомы ДНК выполняется слева направо в направлении $5' \to 3'$ $5' \to 3'$, а у второй комплементарной нити в направлении $5' \to 3'$ справа налево (рис.1, модель Уотсона-Крика). Приводимые ниже соотношения, как правило, выполняются приближенно.

$5'$ _____ AC _____ GT ___ $\rightarrow$ __ $3'$ 1-st string

..... $\updownarrow\updownarrow$ .....

$3'$ _____ TG _____ CA ____ $\leftarrow$ __ $5'$ 2-nd string

Рис 1. Модель Уотсона-Крика

Симметрия оснований. 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 $n(i)$ – количество оснований i, $i \in \{\mathrm{A,C,G,T}\}$, вычисленных на одной нити. Из соотношений (1) вытекает, что количество каждого основания, подсчитанного по первой и второй нити, совпадает:

$$\begin{aligned} n(\mathrm{A},1) &= n(\mathrm{A},2),\ n(\mathrm{T},1) = n(\mathrm{T},2), \\ n(\mathrm{C},1) &= n(\mathrm{C},2),\ n(\mathrm{G},1) = n(\mathrm{G},2) \end{aligned} \tag{2}$$

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

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

или, короче, в виде формулы

$$n(ij) = n(\overline{ji}) \tag{4}$$

где $i, j \in \{\mathrm{A,C,G,T}\}$, $\overline{\mathrm{A}} = \mathrm{T}$, $\overline{\mathrm{C}} = \mathrm{G}$, $\overline{\mathrm{T}} = \mathrm{A}$, $\overline{\mathrm{G}} = \mathrm{C}$. Заметим, что пары AT, TA , CG GC не присутствуют в (3), поскольку они приводят к тавтологии.

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

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

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

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

где $n(ij)$ – число пар (ij), $i, j \in \{\mathrm{A,C,G,T}\}$, $n(i)$ – число оснований i в цепи хромосомы, представляют собой оценки переходных вероятностей для однородных цепей Маркова. Из (5) и (6) вытекает, что вторая комплементарная нить в направлении $5' \rightarrow 3'$ имеет такие же оценки переходных вероятностей $\hat{p}(ij)$, как и исходная первая нить (рис.1).

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

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

где $i \in \{\mathrm{A,C,G,T}\}$, то есть 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),

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

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

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

Утверждение 1. Из симметрии пар оснований вытекает симметрия оснований.

Симметрия троек оснований. Кодоны (тройки оснований) связаны следующими соотношениями:

$$n(ijk) = n(\overline{\overline{k}\,\overline{j}\,\overline{i}}) \quad (10)$$

Здесь $n(ijk)$ – число троек оснований (ijk), $i, j, k \in \{\mathrm{A,C,G,T}\}$, $(\overline{\overline{k}\,\overline{j}\,\overline{i}})$ – антикодон кодона (ijk). Аналогично (5) из соотношений (10) вытекает симметрия относительно записи 64 троек оснований для каждой нити ДНК:

$$n(ijk,1) = n(ijk,2) \quad (11)$$

По аналогии с (8), (9), используя соотношения (10), для пар (3) выводятся шесть связывающих ограничений для троек оснований.

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

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, которая будет подтверждать 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 $(\overline{\overline{k}\,\overline{j}\,\overline{i}})$ совпадают

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

где n – длина хромосомы. Таким образом, ожидаемое число повторов троек оснований (ijk) и $(\overline{\overline{k}\,\overline{j}\,\overline{i}})$ 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 $x_1, x_2, ..., x_{n-1}, x_n$ совпадает с оценкой вероятности последовательности $\overline{x}_n, \overline{x}_{n-1}, ..., \overline{x}_2, \overline{x}_1$, т.е.

$$\hat{p}(x_1, x_2, ..., x_{n-1}, x_n) = \hat{p}(\overline{x_n}, \overline{x_{n-1}}, ..., \overline{x_2}, \overline{x_1}) \quad (12)$$

Отсюда следует, что вероятности двух противоположных нитей хромосомы, подсчитанные в модели однородной цепи Маркова на основе оценок переходных вероятностей (6), совпадают.

2. Одинаковая полярность цепей ДНК

Заметим, что симметрия оснований (1), (2) $n(i,1) = n(i,2)$, $i \in \{\mathrm{A,C,G,T}\}$ может выполняться и в том случае, когда обе комплементарные нити ДНК имеют одинаковые направления записи оснований (рис.2). Однако в природе такой вид симметрии отсутствует.

$5'$ _____ ij _____ $\overline{ij}$ ___ $\rightarrow$ __ $3'$ 1-st string

..... $\updownarrow\updownarrow$ .....

$5'$ _____ $\overline{ij}$ _____ ij ___ $\rightarrow$ __ $3'$ 2-nd string

Рис 2. Одинаковая полярность цепей ДНК

Поэтому из симметрии оснований (1), (2) нельзя вывести симметрию пар оснований. В таком случае симметрия пар оснований $n(ij,1) = n(ij,2)$, $i, j \in \{\mathrm{A,C,G,T}\}$, вытекает из следующих соотношений, выполняющихся для одной нити:

$$n(\mathrm{AA})= n(\mathrm{TT}),\ n(\mathrm{CC})= n(\mathrm{GG}),\ n(\mathrm{AC})= n(\mathrm{TG}),\ n(\mathrm{CA}) = n(\mathrm{GT})$$

$$n(\mathrm{AG})= n(\mathrm{TC}),\ n(\mathrm{CG}) = n(\mathrm{GC}),\ n(\mathrm{AT})= n(\mathrm{TA}),\ n(\mathrm{CT}) = n(\mathrm{GA}) \qquad (13)$$

или в виде одной формулы

$$n(ij) = n(\overline{i}\ \overline{j}) \qquad (14)$$

Заметим, что в отличие от рассмотренной симметрии (4) пары AT, TA , CG, GC, присутствуют в (14), т.е. для пар оснований на два ограничения больше.

Для симметрии (14) ограничения (8), (9) трансформируются в одно ограничение

$$n(\mathrm{CA}) + n(\mathrm{GA}) = n(\mathrm{AC}) + n(\mathrm{AG}) \qquad (15)$$

Симметрия троек оснований $n(ijk,1) = n(ijk,2)$ вытекает из соотношений, выполняющихся для одной нити ДНК:

$$n(ijk) = n(\overline{ijk}) \qquad (16)$$

Для симметрии вида (16) добавляется два связывающих ограничения для троек оснований.

Для симметрии с одинаковой полярностью нитей ДНК справедливы рассмотренные выше утверждения 1–3. Из соотношений (1), (14) вытекает, что для однородной цепи Маркова оценки вероятностей троек оснований (ijk) и $(\overline{ijk})$ совпадают:

$$n\hat{p}(ijk) = \frac{n(i)n(ij)n(jk)}{n(i)n(j)} = n\hat{p}(\overline{ijk}) = \frac{n(\overline{i})n(\overline{ij})n(\overline{jk})}{n(\overline{i})n(\overline{j})}.$$

Утверждение 4 записывается следующим образом.

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

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

Проведенные исследования показывают, что симметрия вида $n(ij) = n(\overline{ji})$ имеет восемь связывающих ограничений (3), (8), (9) для пар оснований, а симметрия $n(ij) = n(\overline{i}\ \overline{j})$ содержит девять ограничений (13), (15). Для троек оснований у симметрии вида $n(ijk) = n(\overline{ijk})$ на два ограничения больше, чем у симметрии $n(ijk) = n(\overline{k}\,\overline{ji})$. Поэтому ДНК с противоположной полярностью нитей имеет больше степеней свободы, чем ДНК с одинаковой полярностью, т.е. с точки зрения теории информации модель Уотсона-Крика более эффективна.

3. Генерация случайных последовательностей с симметриями обеих видов

С помощью модели цепей Маркова можно легко сгенерировать случайные последовательности, для которых будет выполняться симметрия вида (1), (4), (10) для модели Уотсона-Крика. На основе оценок переходных вероятностей (6) и программы псевдослучайных чисел строились случайные последовательности оснований различной длины, в том числе и совпадающие по длине с хромосомами человека. Аналогичным образом генерировались случайные последовательности, для которых выполняются симметрии вида (1), (14), (16) для модели с одинаковой полярностью цепей ДНК. Численные расчеты показали, что относительная разность между тройками оснований в (10) и (16) значительно меньше 1%. Таким образом, на основе модели цепей Маркова генерируются случайные последовательности, для которых выполняется симметрия для двух видов полярности нитей ДНК.

Заключение

Показано, что для нитей ДНК возможно два вида симметрии, но в природе реализована более эффективная с точки зрения теории информации модель Уотсона-Крика. Симметрия отдельных оснований – следствие симметрии пар оснований и соответственно симметрия пар оснований – следствие 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 [2]. Полученные результаты открывают широкие возможности применения байесовских процедур на моделях цепей Маркова для распознавания свойств участков оснований (генов), в том числе генетических заболеваний.

Библиография

1. Гупал А.М., Вагис А.А. Комплементарность оснований в хромосомах ДНК // Проблемы управления и информатики. – 2005. – № 5. – С. 90–94.

2. Гупал А.М., Сергиенко И.В. Оптимальные процедуры распознавания.– Киев: Наукова думка, 2008.– 232 с.

3. Сергиенко И.В., Гупал А.М., Вагис А.А. Правила симметрии в записи генетической информации ДНК // Кибернетика и системный анализ. – 2011. – № 3. – С. 88–94.

Сведения об авторах

Вагис Александра Анатольевна *– Институт кибернетики им. В.М. Глушкова НАН Украины, к.ф.-м.н., с.н.с., Киев, Украина. E-mail: valex_ic@mail.ru*

Гупал Анатолий Михайлович *– Институт кибернетики им. В.М. Глушкова НАН Украины, д.ф.-м.н., чл.-к. НАН Украины, зав. отд., Киев, Украина. E-mail: gupal_anatol@mail.ru*