УДК 576.3.7

# КРИТЕРИЙ ПРАВИЛЬНОСТИ РАМКИ СЧИТЫВАНИЯ ПРИ ТРАНСЛЯЦИИ ДНК

© А.А. Арзамасцев

Arzamastsev A.A. The criterion of the reading frame truth at the DNA translation operation. It is shown that the frequency of arrangement amine-acid chain can present the criterion of the reading frame truth at the DNA translation operation.

**Введение.** Проблема "правильности рамки считывания" существует при трансляции последовательности оснований ДНК в аминокислоты при образовании протеинов [1]. Суть проблемы заключается в следующем. Информация с ДНК (РНК) считывается триплетами (кодонами), каждому из которых соответствует аминокислота или управляющий сигнал типа Start или Stop. Нуклеотидная последовательность может быть считана с помощью любой из трех различных рамок считывания [2]. Рамка считывания устанавливается в момент инициации синтеза полипептидной цепи [3]. При возникновении мутации, заключающейся в сдвиге рамки считывания кодонов на одно или два основания, вся следующая цепочка оснований считывается неверно. Существующие в живых системах механизмы способны с одной стороны обнаруживать неправильную рамку считывания, а с другой - корректировать ее. Однако понимания того, как они решают эти проблемы, в настоящее время нет. Вместе с тем существующие компьютерные алгоритмы анализа ДНК не позволяют решить вопрос о "правильности рамки считывания" из самой последовательности.

**Постановка проблемы.** Необходимо, находясь в любом месте нуклеотидной последовательности, определить правильную рамку ее считывания, не пользуясь информацией о расположении стартовых и терминантных кодонов. Иными словами, требуется определить правильность рамки изнутри, найти какой-либо стационарный признак, существенным образом изменяющийся при изменении рамки считывания.

**Результаты и обсуждение**. При идентификации и анализе текстов, написанных на различных языках, часто используется частота встречаемости букв. Однако при отсутствии информации о самом алфавите (знаковой системе) такой критерий вряд ли может принести пользу, поскольку в этом случае оказывается невозможным установление соответствия между самими знаками и частотой их встречаемости.

Ранее нами было показано, что форма гистограммы ранжированной частоты встречаемости (РЧВ) символов, с одной стороны, инвариантна к самой знаковой системе, а с другой, может являться стационарной характеристикой языка. По этой причине было решено использовать этот критерий при решении проблемы о правильности рамки считывания.

Поскольку исходный алфавит ДНК (РНК) является четырехбуквенным и сдвиг рамки не изменяет частоту встречаемости букв, было решено исследовать на частоту встречаемости не исходный текст, а текст, являющийся результатом трансляции, т. е. аминокислотные последовательности.

Для выполнения работы была разработана программа - транслятор, работающая в соответствии с таблицей генетического кода [2]. Для проверки правильности трансляции сравнивалось действие программы по переводу последовательностей кодонов ДНК в аминокислоты (белки) с истинными последовательностями аминокислот в белках. Для этого использовалась информация о последовательностях нуклеотидов в геноме человека (ген MSH2, 2-я хромосома) и последовательности аминокислот в белке. Результаты трансляции приведены ниже.

**Перевод гена MSH2 с помощью программы-транслятора**

```
                DMAVQPKETLQLESAAEVGFVRFFQGMPEKPTTTVRLFDRGDFYT
AHGEDALLAAREVFKTQGVIKYMGPAGAKNLQSVVLSKMNFESFVKDLLLVRQYRVEV
YKNRAGNKASKENDWYLAYKASPGNLSQFEDILFGNNDMSASIGVVGVKMSAVDGQRQ
VGVGYVDSIQRKLGLCEFPDNDQFSNLEALLIQIGPKECVLPGGETAGDMGKLRQIIQ
RGGILITERKKADFSTKDIYQDLNRLLKGKKGEQMNSAVLPEMENQVAVSSLSAVIKF
LELLSDDSNFGQFELTTFDFSQYMKLDIAAVRALNLFQGSVEDTTGSQSLAALLNKCK
TPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFVEDAELRQTLQEDLLRRFPDLNRLAK
KFQRQAANLQDCYRLYQGINQLPNVIQALEKHEGKHQKLLLAVFVTPLTDLRSDFSKF
QEMIETTLDMDQVENHEFLVKPSFDPNLSELREIMNDLEKKMQSTLISAARDLGLDPG
KQIKLDSSAQFGYYFRVTCKEEKVLRNNKNFSTVDIQKNGVKFTNSKLTSLNEEYTKN
KTEYEEAQDAIVKEIVNISSGYVEPMQTLNDVLAQLDAVVSFAHVSNGAPVPYVRPAI
LEKGQGRIILKASRHACVEVQDEIAFIPNDVYFEKDKQMFHIITGPNMGGKSTYIRQT
GVIVLMAQIGCFVPCESAEVSIVDCILARVGAGDSQLKGVSTFMAEMLETASILRSAT
KDSLIIIDELGRGTSTYDGFGLAWAISEYIATKIGAFCMFATHFHELTALANQIPTVN
NLHVTALTTEETLTMLYQVKKGVCDQSFGIHVAELANFPKHVIECAKQKALELEEFQY
IGESQGYDIMEPAAKKCYLEREQGEKIIQEFLSKVKQMPFTEMSEENITIKLKQLKAE
VIAKNNSFVNEIISRIKVTT{stop}KIPVME{stop}R{stop}Y{stop}{stop}
AIVCNSFILFYINPFSIVLTVSAHGLST{stop}{stop}DI{stop}{stop}YFTL
RTFSKIFILKNESCN{stop}GLFAIDIGNNK{stop}CAEFYK {stop}NHVVC
```

**Истинная последовательность аминокислот протеина, соответствующего гену MSH2 (информация получена с сайта программы "Геном человека" http://www.ncbi.nlm.nih.gov/SCIENCE96/ )**

MAVQPKETLQLESAAEVGFVRFFQGMPEKPTTTVRLFDRGDFYT
AHGEDALLAAREVFKTQGVIKYMGPAGAKNLQSVVLSKMNFESFVKDLLLVRQYRVEV
YKNRAGNKASKENDWYLAYKASPGNLSQFEDILFGNNDMSASIGVVGVKMSAVDGQRQ
VGVGYVDSIQRKLGLCEFPDNDQFSNLEALLIQIGPKECVLPGGETAGDMGKLRQIIQ
RGGILITERKKADFSTKDIYQDLNRLLKGKKGEQMNSAVLPEMENQVAVSSLSAVIKF
LELLSDDSNFGQFELTTFDFSQYMKLDIAAVRALNLFQGSVEDTTGSQSLAALLNKCK
TPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFVEDAELRQTLQEDLLRRFPDLNRLAK
KFQRQAANLQDCYRLYQGINQLPNVIQALEKHEGKHQKLLLAVFVTPLTDLRSDFSKF
QEMIETTLDMDQVENHEFLVKPSFDPNLSELREIMNDLEKKMQSTLISAARDLGLDPG
KQIKLDSSAQFGYYFRVTCKEEKVLRNNKNFSTVDIQKNGVKFTNSKLTSLNEEYTKN
KTEYEEAQDAIVKEIVNISSGYVEPMQTLNDVLAQLDAVVSFAHVSNGAPVPYVRPAI
LEKGQGRIILKASRHACVEVQDEIAFIPNDVYFEKDKQMFHIITGPNMGGKSTYIRQT
GVIVLMAQIGCFVPCESAEVSIVDCILARVGAGDSQLKGVSTFMAEMLETASILRSAT
KDSLIIIDELGRGTSTYDGFGLAWAISEYIATKIGAFCMFATHFHELTALANQIPTVN
NLHVTALTTEETLTMLYQVKKGVCDQSFGIHVAELANFPKHVIECAKQKALELEEFQY
IGESQGYDIMEPAAKKCYLEREQGEKIIQEFLSKVKQMPFTEMSEENITIKLKQLKAE
VIAKNNSFVNEIISRIKVTT

**Перевод с помощью программы-транслятора этого же гена, полученного из базы данных института молекулярной биологии РАН ( Москва)**

MAVQPKETLQLESAAEVGFVRFFQGMPEKPTTTVRLFDRGDFYT
AHGEDALLAAREVFKTQGVIKYMGPAGAKNLQSVVLSKMNFESFVKDLLLVRQYRVEV
YKNRAGNKASKENDWYLAYKASPGNLSQFEDILFGNNDMSASIGVVGVKMSAVDGQRQ
VGVGYVDSIQRKLGLCEFPDNDQFSNLEALLIQIGPKECVLPGGETAGDMGKLRQIIQ
RGGILITERKKADFSTKDIYQDLNRLLKGKKGEQMNSAVLPEMENQVAVSSLSAVIKF
LELLSDDSNFGQFELTTFDFSQYMKLDIAAVRALNLFQGSVEDTTGSQSLAALLNKCK
TPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFVEDAELRQTLQEDLLRRFPDLNRLAK
KFQRQAANLQDCYRLYQGINQLPNVIQALEKHEGKHQKLLLAVFVTPLTDLRSDFSKF
QEMIETTLDMDQVENHEFLVKPSFDPNLSELREIMNDLEKKMQSTLISAARDLGLDPG
KQIKLDSSAQFGYYFRVTCKEEKVLRNNKNFSTVDIQKNGVKFTNSKLTSLNEEYTKN
KTEYEEAQDAIVKEIVNISSGYVEPMQTLNDVLAQLDAVVSFAHVSNGAPVPYVRPAI
LEKGQGRIILKASRHACVEVQDEIAFIPNDVYFEKDKQMFHIITGPNMGGKSTYIRQT
GVIVLMAQIGCFVPCESAEVSIVDCILARVGAGDSQLKGVSTFMAEMLETASILRSAT
KDSLIIIDELGRGTSTYDGFGLAWAISEYIATKIGAFCMFATHFHELTALANQIPTVN
NLHVTALTTEETLTMLYQVKKGVCDQSFGIHVAELANFPKHVIECAKQKALELEEFQY
IGESQGYDIMEPAAKKCYLEREQGEKIIQEFLSKVKQMPFTEMSEENITIKLKQLKAE
VIAKNNSFVNEIISRIKVTT{stop}

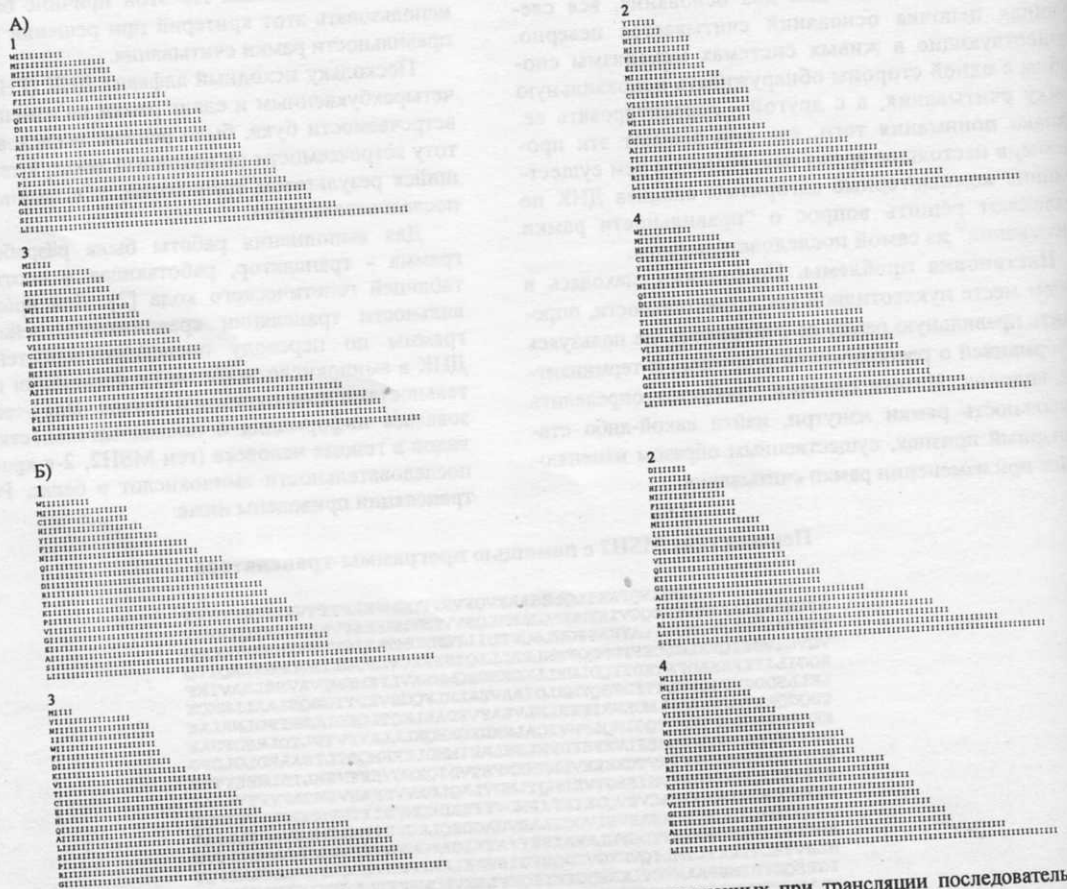


**Рис. 1.** Ранжированные относительные частоты встречаемости аминокислот, полученных при трансляции последовательностей ДНК человека по оси абсцисс. Аминокислоты (номер аминокислот) по оси ординат. А) и Б) - два независимых расчета. 1- правильная рамка считывания, 2- сдвиг на одну букву, 3- сдвиг на две буквы, 4- сдвиг на число букв, кратное трем.

Результаты сравнения позволили убедиться в правильности работы транслятора. Некоторое различие в начале и конце последовательностей обусловлено тем, что здесь приведены результаты трансляции последовательностей с начальными "предстартовыми" и конечными "послеэндными" кодонами. В файлах, полученных из института молекулярной биологии, содержатся лишь кодоны, кодирующие аминокислоты. Данная программа была использована в настоящей работе.

Сдвигом на одну букву будем называть вырезание одной буквы из последовательности оснований ДНК в ее начале, сдвигом на две - вырезание двух букв и т. д. На рис. 1 (А, Б) показаны гистограммы РЧВ букв в аминокислотных последовательностях протеинов, полученные в результате обработки информации о геноме человека, полученной из института молекулярной биологии РАН. Общий объем обработанной информации составляет около 6,7 Mb, что соответствует примерно 150 генам средней длины. Из рис. 1 хорошо видно, что без сдвига рамки считывания или при ее сдвиге на 3 буквы, что фактически соответствует правильной рамке, гистограммы РЧВ имеют вид выпуклых зависимостей. Эта ситуация изменяется при сдвиге на одну и две буквы. На соответствующих гистограммах появляются характерные участки, имеющие вид вогнутых зависимостей.

Таким образом, обработка значительной доли информации из генома человека убедительно показывает, что РЧВ аминокислот в протеиновой последовательности может быть критерием правильности рамки считывания последовательностей оснований ДНК.

Обработка данных по РЧВ, приведенных на рис. 1, по методу наименьших квадратов позволила получить уравнения для расчета относительной величины РЧВ. Без сдвига рамки или со сдвигом на число букв, кратное 3, справедливо уравнение:

$$f(n) = 0{,}014519 + 0{,}000111 \cdot n + 0{,}001031 \cdot n^2 - 0{,}000094 \cdot n^3 + 0{,}000003 \cdot n^4 \qquad (1)$$


**Рис. 2.** Обобщенная блок-схема использования метода.

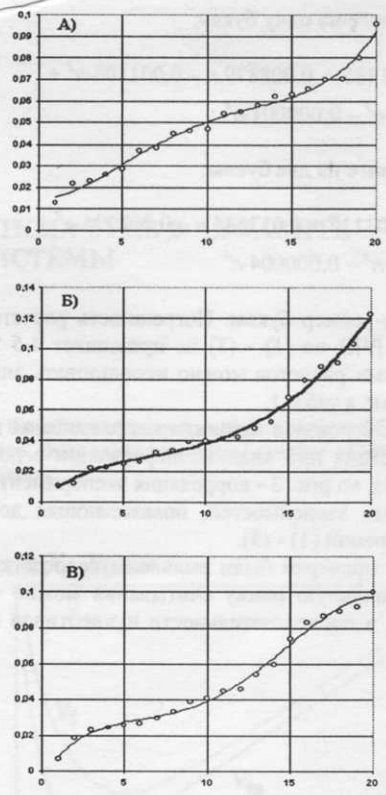


**Рис. 3.** Корреляция экспериментальных данных по ранжированной частоте встречаемости букв с результатами расчетов по уравнениям (1)-(3). По оси абсцисс - $n$, по оси ординат - $f(n)$. А) - сдвиг отсутствует или кратен трем и уравнение (1); Б) - сдвиг равен одной букве и уравнение (2); В) - сдвиг равен двум буквам и уравнение (3).

*Таблица 1.*

Сравнение ранжированных частот встречаемости аминокислот при различных сдвигах рамки считывания оснований ДНК

| № аминокислоты | Ранжированные частоты встречаемости аминокислот в протеиновой последовательности | | |
|---|---|---|---|
| | Без сдвига рамки или со сдвигом на число оснований, кратное трем | Со сдвигом рамки на одно основание | Со сдвигом рамки на два основания |
| 1 | 0,012837916 | 0,011044325 | 0,006497186 |
| 2 | 0,021544848 | 0,012219755 | 0,018819673 |
| 3 | 0,023166387 | 0,022171143 | 0,023285825 |
| 4 | 0,025659831 | 0,022596271 | 0,024355235 |
| 5 | 0,028570225 | 0,025042131 | 0,025839841 |
| 6 | 0,037126843 | 0,025902313 | 0,027100602 |
| 7 | 0,038268667 | 0,032043561 | 0,029693427 |
| 8 | 0,045137062 | 0,033901745 | 0,033331732 |
| 9 | 0,046329149 | 0,03900856 | 0,038753271 |
| 10 | 0,04711925 | 0,039908985 | 0,040845438 |
| 11 | 0,054434937 | 0,040779109 | 0,045235923 |
| 12 | 0,055134806 | 0,04232255 | 0,045963798 |
| 13 | 0,058104575 | 0,047998138 | 0,054361831 |
| 14 | 0,06221734 | 0,052373898 | 0,060194976 |
| 15 | 0,063068484 | 0,068133547 | 0,074227038 |
| 16 | 0,065472131 | 0,079425925 | 0,083259633 |
| 17 | 0,069663076 | 0,088348843 | 0,086763284 |
| 18 | 0,070563352 | 0,091040395 | 0,08953796 |
| 19 | 0,080209301 | 0,103157714 | 0,091859937 |
| 20 | 0,095371818 | 0,122581092 | 0,100073388 |

При сдвиге на одну букву:

$$f(n) = 0{,}001485 + 0{,}008879 \cdot n - 0{,}001108 \cdot n^2 + 0{,}000069 \cdot n^3 - 0{,}000001 \cdot n^4 \quad (2)$$

При сдвиге на две буквы:

$$f(n) = -0{,}003117 + 0{,}013634 \cdot n - 0{,}002296 \cdot n^2 + 0{,}000178 \cdot n^3 - 0{,}000004 \cdot n^4 \quad (3)$$

Здесь $n$- номер буквы. Погрешность расчета относительной РЧВ по (1) - (3) не превышает 4-5 %. Для более точных расчетов можно использовать значения, приведенные в табл. 1.

На рис. 2 показан алгоритм использования предлагаемого метода при анализе информации с помощью компьютера, на рис. 3 - корреляция экспериментальных и расчетных зависимостей, показывающая достоверность уравнений (1) - (3).

В ходе проверки были выявлены недостатки метода. 1) правильную рамку считывания можно определить лишь в последовательности нуклеотидов относительно большой длины (около 6000 оснований); 2) сам живой организм (клетка) не может использовать такой алгоритм или его разновидности для определения правильной рамки.

## ЛИТЕРАТУРА


1. *Инге-Вечтомов С.Г.* Трансляция как способ существования живых систем, или в чем смысл "бессмысленных" кодонов // Соросовский Образовательный Журнал. 1996. № 12. С. 2-10.
2. *Альбертс Б. и др.* Молекулярная биология клетки. М.: Мир, 1986. Т. 1-2.
3. *Hunt T.* The initiation of protein synthesis // Trends Biochem. Sci. 1980. V. 5. P. 178-181.


БЛАГОДАРНОСТИ: Работа выполнена при поддержке Международной Соросовской Программы Образования в Области Точных Наук (*International Soros Science Education Program*) (грант d98-455, 1998). Автор благодарит В.Ю. Макеева (*Институт Молекулярной Биологии РАН, Москва*) за предоставление информации о частично расшифрованном к настоящему моменту геноме человека.

Поступила в редакцию 12 ноября 1998 г.