ОБ ЭКСТРЕМАЛЬНЫХ СВОЙСТВАХ РАЗМЕТКИ ГЕНЕТИЧЕСКОГО КОДА

Науменко С. А., Подлазов А. В.

(Россия, Москва)

Работа посвящена поиску закономерностей в генетическом коде. Показано, что его разметка — разделение кодонов на смысловые и терминирующие — является решением набора оптимизационных задач. Реализовавшаяся разметка кода обеспечивает максимально возможную устойчивость генетической информации по отношению к ошибкам двух классов: чтению со сдвигом и точечным мутациям. Также она наилучшим образом соответствует распространенности в природе простейших органических соединений, какой она была на этапе зарождения жизни.

Введение. Основой жизни являются линейные гетерополимеры — *нуклеиновые кислоты* и *белки*. В состав белков входят 20 видов мономеров — *аминокислот*, а в состав нуклеиновых кислот — 4 вида мономеров — *нуклеотидов*, обозначаемых буквами A, G, C и T.

Аминокислотная последовательность белка определяется нуклеотидной последовательностью кодирующего его гена в соответствии с правилами, называемыми *генетическим кодом*. Одну аминокислоту задает *нуклеотидный триплет*, или *кодон,* — три последовательно идущих нуклеотида. Порядок нуклеотидов значим, поэтому существует $4^3 = 64$ различных кодонов. Из них 61 являются *смысловыми*, кодирующими аминокислоты, а 3 — *терминирующими*, или *стоп-кодонами*, дающими сигнал к прекращению синтеза белка. С формальной точки зрения генетический код есть отображение имеющего определенную внутреннюю

структуру алфавита из 64 триплетов на множество, состоящее из 20 символов (аминокислот) и 1 знака препинания.

Реализовавшийся в природе генетический код является одним из огромного числа возможных, однако он *универсален*, т.е. един для всех организмов (за несколькими незначительными исключениями) [1,2]. Более того, присутствующие в нем закономерности позволяют говорить о неслучайности генетического кода [2,3].

Ранее вопросы о возможных экстремальных свойствах генетического кода исследовались в Институте прикладной математики им. М.В. Келдыша РАН [4]. Эти вопросы естественно возникли в связи с анализом проблем добиологической и ранних стадий биологической эволюции [5].

Настоящая работа посвящена рассмотрению одной из таких закономерностей, связанной с *разметкой кода* — делением кодонов на смысловые и терминирующие. Разметку универсального кода будем называть *канонической*. Для нее терминирующими являются триплеты TAA, TAG и TGA.

Здесь мы ограничимся разметками с 3 стоп-кодонами, как у канонической разметки. Всего существует $C_{64}^{3} = 41\,664$ варианта таких разметок.

Чтение со сдвигом. Признаки начала и конца нуклеотидного триплета в геноме отсутствуют. Чтение нуклеотидной цепи происходит последовательно, триплет за триплетом. Поэтому существуют три принципиально разных способа прочесть одну и ту же нуклеотидную последовательность, которые определяются *рамкой считывания* — положением первого нуклеотида при чтении (рис. 1.).

В подавляющем большинстве случаев функциональный белок синтезируется только при одной рамке считывания. Более того, чтение нуклеотидной последовательности с неправильным положением рамки часто оказывается заблокированным появле-

нием терминирующего кодона, который делает недоступным всю расположенную за ним информацию (рис. 1.)

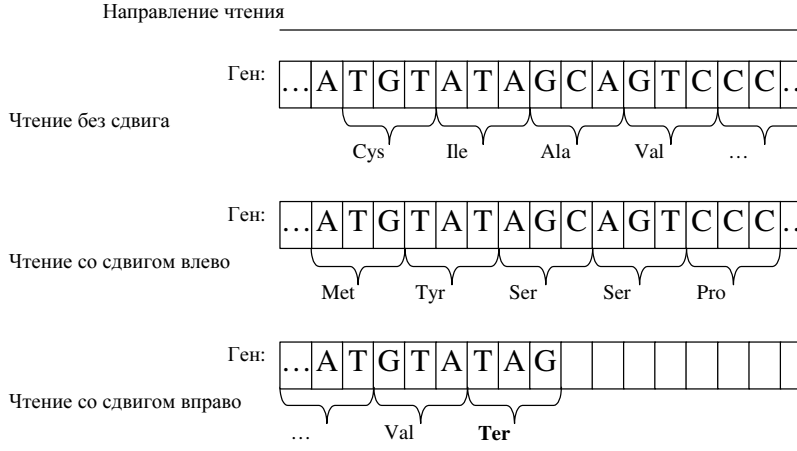

Рис. 1. Примеры чтения гена.
Нуклеотиды в гене группируются в триплеты, кодирующие последовательность аминокислот. При смещении рамки считывания набор триплетов оказывается совершенно иным.
Вверху — чтение без сдвига. Ген кодирует аминокислотную последовательность Cys–Ile–Ala-Val–…
По центру — чтение со сдвигом на один нуклеотид влево. Тот же ген кодирует совсем другую аминокислотную последовательность: Met–Tyr–Ser–Ser–Pro–…
Внизу — чтение со сдвигом на один нуклеотид вправо. Последовательность смысловых кодонов прервана терминирующим кодоном (Ter). Дальнейшее чтение гена невозможно

Пару смысловых кодонов, которая при чтении со сдвигом дает терминирующий кодон, будем называть *запретной*. Такова, например, пара A**TA–G**CA, при чтении которой со сдвигом на один нуклеотид вправо возникает терминирующий триплет TAG. Пара идущих подряд аминокислот будет запретной лишь в том

случае, если для нее все возможные пары задающих ее кодонов запретны. Например, пара кодонов ATA+GCA запретна, но задаваемая ею пара аминокислот Ile+Ala – нет, т.к. ее же можно закодировать и сочетанием ATT+GCA, не приводящим к появлению терминирующего кодона при чтении с иной рамкой. А вот пара аминокислот Met+Lys является запретной, т.к. любая пара соответствующих ей кодонов (ATG+AAA или ATG+AAG) при чтении со сдвигом вправо порождает стоп-кодон. В теореме [4] речь идет именно о запретных парах аминокислот, для нас же, поскольку мы ограничиваемся разметкой кода, значение имеют только запретные пары кодонов.

Мутации. Кодоны, различающиеся между собой лишь одним нуклеотидом в одном и том же положении (т.е. на первом, втором или третьем месте в триплете), будем называть *соседними*. Соседние кодоны могут быть получены друг из друга в результате *точечных мутаций* — случайной замены одного нуклеотида другим из-за ошибок в процессе воспроизводства или прочтения генома. Точечные мутации применительно к (смысловым) кодонам подразделяются на *миссенс-мутации* и *нонсенс-мутации* в зависимости от того, превращается ли кодон в результате в другой смысловой кодон или становится терминирующим.

С биологической точки зрения миссенс-мутации, локально изменяющие закодированную аминокислотную последовательность, не очень опасны по сравнению с нонсенс-мутациями, обрывающими ее и делающими тем самым невозможным получение функционального белка. Поэтому смысловые кодоны, соседствующие с терминирующими и вследствие этого подверженные нонсенс-мутациям, будем называть *уязвимыми*.

Постановка задач оптимизации. Для любого кода решающее значение при передаче данных имеет устойчивость к ошибкам. В случае генетического кода можно выделить два основных источника ошибок, связанных с разметкой: смещение рамки считывания и точечные мутации. Задачи обеспечения ус-

тойчивости по отношению к ошибкам этих классов принципиально различны.

Задача I — блокировка чтения со сдвигом. Смещение рамки считывания — это нелокальная ошибка, полностью искажающая генетическую информацию. Обеспечить ее восстановление в этом случае невозможно, и единственный выход — как можно скорее пресечь бессмысленную работу по чтению несуществующего гена. Поэтому оптимальной является разметка кода, обеспечивающая максимальную вероятность появления запретных пар кодонов при чтении со сдвигом.

Задача II — обеспечение устойчивости к точечным мутациям. Случайная замена одного нуклеотида на другой является локальной ошибкой, последствия которой возможно уменьшить. С этой точки зрения оптимальна разметка, минимизирующая вероятность нонсенс-мутаций.

Задача II распадается на 2 подзадачи. Подзадача II-a формулируется как задача минимизации суммарного (по всему коду) числа возможных нонсенс-мутаций, а подзадача II-b — как минимизация числа уязвимых кодонов.

Решение задач оптимизации. Задача I. Существует $61 \times 61 = 3\,721$ пара смысловых кодонов — последовательностей из шести букв из набора {A;G;C;T}, ни одна из трехбуквенных половин которых, стоящих в положениях 1–2–3 и 4–5–6, стоп-кодоном не является. Пара смысловых кодонов запретна, если она содержит три подряд идущих буквы, образующих стоп-кодон, в положениях 2–3–4 или 3–4–5.

Очевидно, максимально возможное число запретных пар для каждого типа сдвига рамки есть $192 = 3 \times 4^3$ (3 варианта стоп-кодона $\times$ 4^3 варианта остальных нуклеотидов). Причем максимум может достигаться для обоих типов сдвига только одновременно. В самом деле, необходимым и достаточным условием оптимальности разметки в смысле задачи I является отсутствие таких стоп-кодонов, что конец одного из них является продолжением другого (например, TA**G** и **G**CT или T**AG** и **AG**C) или самого себя (на-

пример, **CCC** или **GTG)**. Легко убедиться, что каноническая разметка удовлетворяет этому условию, т.е. является оптимальной, давая 192 запретные пары кодонов в случае обоих сдвигов рамки.

Полный перебор вариантов показывает, что всего существует 2 432 (5,8%) таких разметок. Гистограмма распределения разметок по количеству запретных пар кодонов для сдвига влево представлена на (рис. 2.) Распределение для сдвига вправо выглядит точно так же, причем ни для одной разметки число запретных пар для сдвигов влево и вправо не отличается более чем на 3.

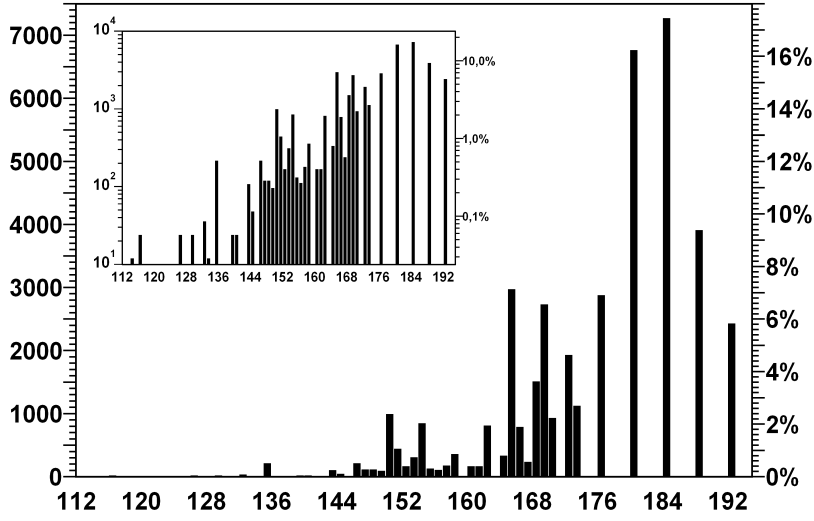

Рис. 2. Распределение разметок по числу запретных пар кодонов.
По оси абсцисс отложено количество запретных пар кодонов, по осям ординат — количество разметок с данным значением числа запретных пар и их доля от общего числа.
На врезке — те же данные с логарифмическим представлением по оси ординат

Решение задач оптимизации. Задача II. С точки зрения устойчивости к точечным мутациям разметки можно разбить на 4 группы, определяемые взаимным соседством стоп-кодонов.

a) Каждый стоп-кодон является соседним по отношению к двум другим. В этом случае 1 смысловой кодон является соседом сразу всех трех стоп-кодонов и еще по 6 смысловых кодонов соседствуют с каждым из них (2 положения в триплете, по которым совпадают стоп-кодоны, × 3 основания, отличных от стоящих в этих положениях). Итого получается 1+6+6+6 = 19 уязвимых кодонов и 1×3+6+6+6 = 21 возможная нонсенс-мутация. Последнее значение является минимально возможным, т.е. обеспечивает оптимум в смысле задачи II-a.

Всего существует 27 мутаций, превращающих некие кодоны в терминирующие (3 стоп-кодона × 3 положения в триплете × ×3 основания, на которые заменяется основание в мутирующем кодоне). Каждая точечная мутация, превращающая один стоп-кодон в другой, уменьшает число мутаций, идущих на то, чтобы превращать в них смысловые кодоны. Количество мутаций, переводящих стоп-кодоны друг в друга, не может быть более 6, в данном случае их именно столько.

Всего эта группа содержит 192 разметки или 0,46% от общего числа.

b) Один стоп-кодон соседствует с двумя другими, но те, в свою очередь, друг другу соседями уже не являются. Именно такова каноническая разметка (стоп-кодоны: TAA, TAG и TGA — первый соседствует с остальными). Мутаций, переводящих стоп-кодоны друг в друга, здесь уже только 4, соответственно на смысловую часть кода остаются 23 нонсенс-мутации.

В этом случае найдется 5 смысловых кодонов, соседствующих сразу с двумя стоп-кодонами (TAC, TAT, TCA, TTA, TGG) и еще 23−5×2 = 13, которые являются соседями только одного стоп-кодона (AAA, AAG, AGA, GAA, GAG, GGA, CAA, CAG, CGA, TCG, TTA, TCG, TTA). Итого: 5+13 = 18 уязвимый кодон, что, как показывает анализ всех вариантов, является минимально

возможным значением, обеспечивающим оптимум в смысле задачи II-b.

Всего эта группа содержит 1 728 разметок или 4,15% от общего числа.

c) Два стоп-кодона являются соседями, а третий не соседствует ни с одним из них. Мутаций, переводящих стоп-кодоны друг в друга, остается всего 2, что дает 25 нонсенс-мутаций. Число уязвимых кодонов варьируется от 20 до 23.

d) Соседних кодонов среди стоп-кодонов нет. Все 27 нонсенс-мутаций приходятся на смысловую часть кода. Число уязвимых кодонов — от 21 до 27.

Группы разметок c и d мы подробно не рассматриваем, т.к. они не обеспечивают оптимальности в смысле задачи II.

Отметим тот факт, что природа сочла решение задачи II-b более важным, чем решение задачи II-a, а это открывает широкое поле для гипотез и интерпретаций.

Многокритериальная оптимизация. Каноническая разметка является решением и задачи I, и задачи II-b, для которых имеется соответственно 1 728 и 2 432 оптимальные разметки. Если же поставить задачу оптимизации по обоим критериям, то количество возможных решений уменьшится до 528, что составляет менее 1,3% всех возможных разметок.

Особенности абиогенного синтеза органических молекул на этапе зарождения жизни, скорее всего, должны были привести к преимущественному использованию нуклеотидов A и T в геноме [5] и в т.ч. в такой его важной части, как стоп-кодоны [6]. Количество букв из набора {A; T}, входящих в их состав, для канонической разметки равно 7. Это наибольшее значение, при котором еще возможно удовлетворить критериям задачи I.

Если сформулировать максимизацию использования букв A и T в стоп-кодонах как третий — эволюционно-химический — критерий оптимизации, то с его учетом количество оптимальных разметок сокращается до 40 или менее 0,10% их общего числа [6].

Выводы. Каноническая разметка генетического кода обеспечивает максимально возможную устойчивость структуры генетической информации по отношению к ошибкам двух классов: чтению со сдвигом и точечным мутациям. Также она наилучшим образом соответствует распространенности в природе простейших органических соединений, какой она была на стадии зарождения жизни.

Два информационных и один эволюционно-химический критерий ограничивают множество оптимальных разметок менее чем одной тысячной от их общего числа.

СПИСОК ЛИТЕРАТУРЫ

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. Фриленд С., Херст Л. Закодированная эволюция// В мире науки,. 2004. — №7.
3. Freeland S.J., Hurst L.D. The genetic code is one in a million// J. Mol. Evol. — 1998. — N47. — P.238–248.
4. Козлов Н.Н. Теорема для генетического кода// ДАН, 2002. — Т.382. — №5. — С. 593–597.
5. Галимов Э.М. Феномен жизни: Между равновесием и нелинейностью. Происхождение и принципы эволюции. — М.: Эдиториал УРСС, 2001. — 256 с.
6. Науменко С.А., Подлазов А.В. Об экстремальных свойствах генетического кода// ИПМ им. М.В. Келдыша РАН, 2005. — Препринт №129.

ON EXTREME PROPERTIES OF THE GENETIC CODE MARKUP

Naumenko S. A., Podlazov A. V.

(Russia, Moscow)

This article is dedicated to the search of genetic code regularities. It is shown, that the genetic code markup — the codon set separation into semantic and terminator groups of codons — is the solution of the set of optimization problems. Existing genetic code markup provides the maximum possible stability of genetic information with respect to two classes of fault: reading with offset and point mutations. Also it has the best possible correpondense with the simplest organic compounds prevalence in nature on the origin of life stage.