

УДК 519.72

ИНТЕГРАЛЬНЫЕ ХАРАКТЕРИСТИКИ ГЕНЕТИЧЕСКОГО КОДА

© 2010 г. *Н.Н. Козлов*

Институт прикладной математики им. М.В. Келдыша РАН, Москва
e-mail: gencod@keldysh.ru

Работа выполнена при финансовой поддержке Программы фундаментальных исследований Президиума РАН «Интеллектуальные информационные технологии, математическое моделирование, системный анализ и автоматизация», Российского фонда фундаментальных исследований (коды проектов 08-01-00042, 10-01-00145), а также гранта ведущих научных школ (НШ-6700.2010.1).

Гигантские по объему данные, полученные для полных последовательностей ДНК больших геномов (человека и др.), поставили перед биоматематиками задачу их изучения. Данная работа представляет собой теоретические основы для решения ряда задач и прежде всего для расчетов больших наборов генов. Речь идет о т.н. природной блокировке генов, когда все 5 последовательностей кодонов, альтернативных последовательности гена, содержат многократные остановки синтеза белка. Теорема 1 устанавливает потенциал такой блокировки для стандартного генетического кода. Теорема 2 устанавливает потенциал кода, который используется для особых записей генетической информации – т.н. перекрывающихся генов. Устанавливается взаимосвязь между интегральными характеристиками генетического кода, которые могут быть вычислены на основе указанных теорем.

Ключевые слова: генетический код, блокировка генов, перекрывающиеся гены, большой геном, геном человека.

INTEGRAL CHARACTERISTICS OF THE GENETIC CODE

N.N. Kozlov

Keldysh Institute of Applied Mathematics of Russian Academy of Sciences, Moscow

Huge on volume the data received for complete sequences DNA large genomes (of the human etc.), have put before biomathematics a task of their study. The given work represents theoretical bases for the decision of a number of tasks and first of all for accounts of the large sets of genes. The speech goes about so-called natural blocking of genes, when all 5 sequences codons, the alternative sequences of a gene, contain set stops of synthesis of protein. The theorem 1 establishes potential of such blocking for a standard genetic code. The theorem 2 establishes potential of a code, which is used for the singular of the writing genetic information - overlapping genes. The interrelation between the integrated characteristics of a genetic code is established which can be calculated on the basis of the specified theorems.

Key words: genetic code, blocking genes overlapping genes, large genome, human genome.

Современное развитие биоматематики характеризуется новой ситуацией, возникшей после 2003 года, когда был расшифрован геном человека. Этот этап характеризуется наличием огромных по объему, полных последовательностей ДНК для уже десятков организмов (данные на начало 2010 г.). Идет непрерывное нарастание этих данных, которые сразу же выкладываются в Интернете. Возникает новая задача, связанная с многосторонним, в том числе математическим, анализом подобной громадной информации. При этом может быть поставлен ряд совершенно новых математических задач. Некоторые из таких задач нами уже решены, и полученные результаты готовятся к публикации в данном журнале. Однако прежде чем представлять эти результаты, необходимо обратиться к теоретическому исследованию таких задач, о котором речь пойдет ниже, и на результатах которых основаны расчеты по большим геномам. Наши постановки задач обусловлены логикой предыдущего этапа исследований, кратко представленного в статьях данного журнала [1-3]. Они отражают всего лишь один подход к указанному изучению больших геномов. Речь идет о математическом анализе блокировки генов, которая, как полагают биологи, наблюдается для подавляющего большинства генов [4]. Ниже формулируется теорема 1, на основе которой устанавливается потенциал генетического кода, который может использовать природа для генной блокировки. Основные используемые биологические термины представлены в [5].

1. Математический анализ структурных генов

Структурными генами называют гены, кодирующие белковые последовательности. Каждая из таких последовательностей состоит из аминокислот (amino acide) – *aa*. Число *aa* равно 20, и каждая из них кодируется различающимися кодонами. Любой кодон или триплет состоит из трех букв четырехбуквенного алфавита: А, Т, С, G – это нуклеотиды аденин, тимин, цитозин и гуанин соответственно. Таким алфавитом записаны ДНК всех живых организмов, на одной из двух цепей которой располагается структурный ген, записанный последовательностью указанных букв без всяких пропусков, запятых и т.п. Экспериментально установленная таблица стандартного генетического кода показала, что каждая из *aa* кодируется одним, двумя, тремя, четырьмя, либо шестью кодонами (см. табл.1 из [1]). Помимо кодировок *aa* были установлены три кодона

$$\text{ter: TAA, TAG, TGA,} \quad (1)$$

являющиеся кодонами терминации – *ter*, на которых синтез белка останавливается. Если в результате мутации в структурном гене образуется один из трех кодонов *ter*, это приводит к преждевременному останову синтеза белка в том месте, где расположен мутантный кодон. При этом нарушится функция белка, т.к. синтезируется только часть белковой молекулы.

Вследствие трехбуквенной кодировки *aa*, одному и тому же гену соответствует 6 рамок считывания (РС) или 6 последовательностей триплетов по три РС, читающихся в разных направлениях для каждой из двух цепей ДНК. Изучают два вида РС [1]: открытые – ОРС и заблокированные – БРС. Их различие в том, что ОРС не содержит кодонов *ter* из (1), за исключением одного, который располагается в конце гена. Для определенности обозначим через РС0 рамку считывания, соответствующую гену, РС1 и РС2 – две РС, сдвинутые относительно РС0 на –1 и +1 нуклеотид соответственно и относящиеся к той же цепи ДНК, что РС0. К другой цепи ДНК отнесем РС3, РС4, РС5, т.е. РС, сдвинутые

относительно PC0 на $-1, 0, +1$ нуклеотид соответственно. Положим, что PC0-PC2 читаются слева направо, PC3-PC5 – справа налево. В [4] указывается, что для подавляющего большинства генов все 5 альтернативных PC1-PC5 являются БРС. Биологи говорят о мощной биологической защите, которая состоит в том, что если за счет мутаций от PC0 перейдем на одну из пяти PC (PC1-PC5), то белок (как правило, эфемерный) синтезируется лишь частично, что не должно повлиять на нормальное функционирование клетки. На рис.1 представлен начальный участок гена и все альтернативные PC. Кодоны *ter* из (1) обозначены *. В рис.1 используется стандартное трехбуквенное сокращение каждой из 20 *aa* набора A^0 .

A^0 :Phe,Тур,His,Asn,Asp,Cys,Ile,Val,Pro,Thr,Ala,Gly,Ser,Leu,Arg,Met,Trp,Gln,Lys,Glu. (2)

Основная задача состоит в том, чтобы установить потенциал генетического кода, который использует природа для указанной блокировки. В формулировке основного результата ограничимся рассмотрением только блокировок, образующихся парами *aa*. Это вовсе не значит, что число блокирующих кодонов *ter* в каждом конкретном гене является суммой *ter* всех пар *aa* в гене. Можно показать, что такое число зависит также от некоторых сочетаний следующих подряд трех *aa*, четырех *aa* и т.д.

```

OPC0           MetSerIleLysLeuSerTyrArgGluSerPheSerIleLeuGluGluVal...
BPC1 (-1) TyrGluHis * Thr * Leu * ArgValIle * TyrIleArgGlyGly...
BPC2 (+1)      * AlaLeuAsnLeuValIleGluSerHisLeuValTyr * ArgArgPhe...
→             TATGAGCATTAACCTTAGTTATAGAGAGTCATTTAGTATATTAGAGGAGGTTTA...
←             ATACTCGTAATTTGAATCAATATCTCTCAGTAAATCATATAATCTCTCCAAAT...
BPC3 (-1) IleLeuMetLeuSerLeu * LeuSerAspAsnLeuIleAsnSerSerThr ...
BPC4 (0)      HisAlaAsnPheLysThrIleSerLeu * LysThrTyr * LeuLeuAsn...
BPC5 (+1)     SerCys * Val * AsnTyrLeuThrMet * TyrIleLeuProProLys...
```

Рис.1 Шесть PC для фрагмента гена (начало с кодона ATG(Met), направление чтения указано стрелкой →, из которых одна PC является открытой – OPC0 (в ней 17 смысловых кодонов или кодонов, кодирующих только аминокислоты – *aa*), а 5 PC – альтернативные PC – являются блокированными: БРС1-БРС5. При этом БРС3-БРС5 соответствуют другой цепи ДНК и чтение последовательностей кодонов осуществляется в обратном направлении (←). В скобках указан сдвиг в нуклеотидах относительно OPC0. Символом * обозначены каждый из трех кодонов *ter* из (1). Такие кодоны, как показывает анализ природных генов, многократно присутствуют в каждой из пяти БРС1-БРС5.

Примем обозначение для пар *aa*, блокирующих PC, с номером *k*

$$A_k^1 A_k^2, \quad k = 1,2,3,4,5. \quad (3)$$

A_k^1, A_k^2 – единичные aa из (2) либо их наборы, верхний индекс указывает порядок следования aa в паре: первая aa – это 1, вторая – 2.

Имеет место

Теорема 1. Блокировки PC1-PC5 могут создаваться соответственно:

1) PC1 – 90 парами aa :

$$A_1^1, A_1^2, \quad (4)$$

где A_1^1 есть набор 15 первых aa из набора A^0 из (2),

A_1^2 : Asn, Lys, Asp, Glu, Ser, Arg;

2) PC2 – 43 парами aa :

$$\text{Met}A_2^2 \cup A_2^1\hat{A}_2^2, \quad (5)$$

где A_2^2 : Met, Asn, Lys, Ile, Thr, Ser, Arg,

A_2^1 : Ile, Val, Leu,

\hat{A}_2^2 : $A_2^2 \cup \tilde{A}_2^2$, \tilde{A}_2^2 : Asp, Glu, Val, Ala, Gly;

3) PC3 – 45 парами aa :

$$A_3^1A_3^2, \quad (6)$$

где $A_3^1 = A_1^1$, A_3^2 : Tyr, His, Gln;

4) PC4 – двумя aa :

$$\text{Ser, Leu}; \quad (7)$$

5) PC5 – 56 парами aa :

$$A_5^1A_5^2, \quad (8)$$

где A_5^1 : Phe, Ile, Val, Leu, Pro, Thr, Ala, Ser,

$A_5^2 = A_2^2$.

В формулировке теоремы во всех блокировках указаны пары aa , кроме (7), где приведены две aa . На рис.2а представлены все эти пары aa : по оси абсцисс дана первая aa из соответствующей пары, по оси ординат – вторая aa . Оказалось, что суммарное число 234 пар aa , фигурирующих в (4)-(6), (8), свелось к 175 различающимся парам. Эти пары можно разделить по две группы: пары aa , по которым возможна блокировка лишь одной PC, их всего 125, они обозначены цифрами 1-5, указывающими номер блокирующей PC. От PC1 до PC5 соответственно. Оставшиеся 50 пар aa могут быть использованы в блокировках от 2 до 4 PC, но, как показал анализ, одновременно могут блокировать не более двух PC. На рис.2а они представлены двухзначными числами. Причем эти числа

могут состоять из цифр от 1 до 5, что указывает на номера соответствующих РС, так и цифр вне этого диапазона (см. подпись к этому рис.). Отметим, что среднее значение числа блокирующих пар *aa* около 9. Минимум – 0 – соответствует *aa* Trp (триптофан), как оказалось он не участвует в блокирующих парах ни как первая *aa*, ни как вторая *aa* в паре. Кроме того, также три *aa*: Gln, Lys, Glu не участвуют в качестве первой *aa* в блокирующих парах, а 4 *aa*: Phe, Cys, Pro, Leu – в качестве второй *aa* в парах.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Σ
1 Met	2		5									25	25	5	5	5		5	65		9
2 Trp																					0
3 Phe																					0
4 Tyr			3	3	3	3	3	3				3	3	3	3	3	3	3	3	3	15
5 His			3	3	3	3	3	3				3	3	3	3	3	3	3	3	3	15
6 Asn	2		15	1	1	1	1	1				17	17	15	15	15	1	15	75	1	16
7 Asp			1	1	1	1	1	1				12	12	1	1	1	1	1	16	1	15
8 Cys																					0
9 Gln			3	3	3	3	3	3				3	3	3	3	3	3	3	3	3	15
10 Lys	2		15	1	1	1	1	1				17	17	15	15	15	1	15	75	1	16
11 Glu			1	1	1	1	1	1				12	12	1	1	1	1	1	16	1	15
12 Ile	2		5									25	25	5	5	5		5	65		9
13 Val												2	2						24		3
14 Pro																					0
15 Thr	2		5									25	25	5	5	5		5	65		9
16 Ala												2	2						24		3
17 Gly												2	2						24		3
18 Ser	2		15	1	1	1	1	1				17	17	15	15	15	1	15	75	1	16
19 Leu																					0
20 Arg	2		15	1	1	1	1	1				17	17	15	15	15	1	15	75	1	16
Σ	7	0	12	9	9	9	9	9	0	0	0	15	15	12	12	12	9	12	15	9	175

Рис.2а. Представлены пары аминокислот – *aa*, блокирующие РС1-РС5. По оси абсцисс указывается первая *aa* из пары, по оси ординат – вторая. Кроме одиночных цифр (1-5) в клетке (это блокировки одной РС: РС1-РС5 соответственно), указаны также 50 пар *aa*, которые могут соответствовать разным РС, от 2 до 4. Двум РС соответствуют 33 пары *aa*: 4 пары 12 (блокировка РС1, РС2), 20 пар 15(-РС1, РС5), 3 пары 24(-РС2, РС4), 6 пар 25 (-РС2, РС5). Трех и более РС соответствуют 17 пар: 2 пары 16 (-РС1, РС2, РС4), 8 пар 17 (-РС1, РС2, РС5), 3 пары 65 (-РС2, РС4, РС5), 4 пары 75 (-РС1, РС2, РС4, РС5). Все клетки соответствуют 175 парам *aa* (4)-(6), (8). Включение одиночных *aa* из (7) (см. (7')) приводит к расширению этого множества пар до 210.

Разобьем набор 175 пар на два множества пар. Первое из них соответствует т.н. неизбежным блокировкам: это 31 пара *aa*, на рис.2а они выделены жирным шрифтом. Можно видеть, что для этих пар могут блокироваться все РС от РС1 до РС5. В левой части рис.2б дается пример такой блокировки. Так, для пары ValLys в клетке указано число 17, что указывает блокировки для РС1, РС2, РС5. Для любого N: А, С, Т, G имеет место блокировка ТАА (при N: А), TGA в РС5 (при T: С), TGA в РС2 (при T: G) либо дважды ТАА одновременно в РС1 и РС5 (при N: Т).

V a l L y s G T N A A X		V a l G l u G T N G A X
N: A - G T A A A X →		N: A - G T A G A X →
N: C - G T C A A X A G T ←		N: C - G T C G A X C A G C T Y
N: G - G T G A A X →		N: G - G T G G A X C A C C T Y
N: T - G T T A A X → C A A T T Y ←		N: T - G T T G A X → C A A C T Y

Рис.26. Неизбежные блокировки для пары *aa* ValLys (слева) и возможные блокировки для пары *aa* ValGlu (справа). Кодоны *ter* из (1) заштрихованы. Используются кодировки для стандартного кода Val: GTN, Lys: AAX, Glu: GAX, где N: A, C, T, G; X: A, G; Y: T, C. Стрелки указывают направление чтения.

Эта ситуация существенно отличается, скажем, от пары ValGlu (правый рис.26), где блокировки возникают лишь для PC2 (при N: A) и PC1 (при N: A). (В клетке для ValGlu указано число 12 или блокироваться могут лишь PC1 и PC2). При N: G, C блокировки не возникают. Подобные пары отнесем к множеству возможных блокировок. Для поиска полного числа этого множества обратимся к блокировкам (7) (они могут давать блокировки PC4 лишь для ограниченных кодировок Ser и Leu), которые также целесообразно свести к парам *aa*. С этой целью вводятся пары *aa*:

$$A^0 \text{ Ser, Ser } A^0; \quad A^0 \text{ Leu, Leu } A^0, \quad (7')$$

где A^0 – полный набор (2) из 20 *aa*. Включение 80 пар *aa* из (7') в указанное выше множество 175 пар получим лишь 210 различающихся пар *aa*.

Тем самым нами получена двухкомпонентная интегральная характеристика стандартного генетического кода, в которой из полного числа 400 пар установлено минимальное число q_{min} блокирующих пар и их максимальное число q_{max} . Имеем

$$q: \quad q_{min}=31, \quad q_{max}=210. \quad (9)$$

Т.о., оказалось, что в процессе блокировки из 400 возможных пар могут участвовать не более чем 210, а оставшиеся 190 пар не имеют никакого отношения к блокировкам и кодировки в этих парах используются для решения некоторых других задач, которые в настоящее время нами изучаются. Еще раз подчеркнем, что стандартный код устроен так, что независимо от кодировок в 31 паре *aa*, которые соответствуют q_{min} , неизбежно возникают блокировки. Или для гена, который образован из равновероятного набора пар *aa* из всех 400 неизбежно блокируются $31/400 \sim 0.08$, или в среднем возникает (в каких-либо PC1 – PC5) по 8 кодонов *ter* из (1) на каждые 100 смысловых кодонов гена.

Наиболее важным было понять, как же на самом деле природа блокирует гены, используя интегральную характеристику q из (9). В связи с этим, на основе теоремы 1 и ее

следствий были проведены расчеты природных блокировок для 12 больших геномов (более 200 000 генов), в т.ч. для всех известных на сегодня генов генома человека. Результаты исследования готовятся к печати.

2. Математический анализ перекрывающихся генов

Ниже анализируются только пары генов, взаимозависимые между собой. Такие гены называются перекрывающимися и впервые были открыты в 1976 году [1]. Тогда впервые был экспериментально обнаружен случай кодировки двух негомологичных белков, гены которых были записаны со сдвигом на один нуклеотид. К настоящему времени уже установлены все возможные случаи перекрывания пар генов; таких случаев всего 5 (рис.3), причем в двух из них (верхняя полоса на рис.1) участвует только одна цепь ДНК, а в трех других (нижняя полоса на рис.1) – обе цепи ДНК. Теорема 2, которая приводится ниже, устанавливает все возможности по формированию полного набора перекрытий, которые создаются структурой кода в каждом из названных пяти случаев.

ПЕРЕКРЫТИЯ ГЕНОВ ИЗ ОДНОЙ ЦЕПИ ДНК

сдвиг - 1	сдвиг + 1
$B_{11} \rightarrow$ MetGlyAsn... $B_{12} \rightarrow$ AsnGlyGlnGln... ...AATGGGCAACA... <div style="text-align: center;">1</div>	$B_{21} \rightarrow$ MetAlaAla... $B_{22} \rightarrow$...ProTrpLeuLeu... ...CCATGGCTGCTC... <div style="text-align: center;">2</div>

ПЕРЕКРЫТИЯ ГЕНОВ ИЗ РАЗНЫХ ЦЕПЕЙ ДНК

сдвиг - 1	сдвиг 0	сдвиг + 1
$B_{31} \rightarrow$...HisGlyArg... ...CCACGGACGC... ...GGTGCCTGCGCA... ...TrpProArgThr... ← B_{32} <div style="text-align: center;">3</div>	$B_{41} \rightarrow$ MetGluAsn ...ATGGAGAAT... ...TACCTCTTA... ...HisLeuIle... ← B_{42} <div style="text-align: center;">4</div>	$B_{51} \rightarrow$...AsnPheHisGlu... ...AATTTCCACGAG... ...TTAAAGGTGCTC... ...AsnGlyArg... ← B_{52} <div style="text-align: center;">5</div>

Рис.3. Пять возможных случаев перекрываемости генов, соответствующих одной (1,2) либо двум цепям ДНК (3-5). Чтение текстов при этом осуществляется в разных направлениях (указано стрелкой): слева направо для B_{11} , B_{12} , B_{21} , B_{22} , B_{31} , B_{41} , B_{51} и справа налево для B_{32} , B_{42} , B_{52} .

Запись аминокислотных последовательностей осуществляется на основе генетического кода по тексту гена и приводится над таким текстом для плюс-цепи ДНК и под таким текстом для минус-цепи. Причем из-за антипараллельности цепей ДНК чтение этих последовательностей происходит слева направо для плюс-цепи и справа налево для минус-цепи (см. стрелку на рис.3). Сдвиги, указанные на рис.3, обозначают число нуклеотидов, на которые сдвинуты соответствующие гены: для перекрытий в одной цепи ДНК: ген B_{12} сдвинут на -1 нуклеотид по отношению к гену B_{11} , а ген B_{22} – на +1 нуклеотид по отношению к гену B_{21} . На рис.3 приведены также 3 случая перекрытий пар генов из

разных цепей ДНК: сдвиг -1 (B_{32} относительно B_{31}), сдвиг 0 (B_{42} относительно B_{41}) и сдвиг $+1$ (B_{52} относительно B_{51}). Можно видеть, что случаи перекрытий отвечают случаям, когда соответствующие РС1-РС5 из рис.1 становятся последовательно ОРС (вместо БРС) или не содержат кодонов терминации.

Исследования [1-3, 6-13] по математическому анализу перекрывающихся генов относятся только к случаям 1, 2 из рис.3, т.е. к перекрытиям в одной цепи ДНК. На основе изучения экспериментальных данных по таким перекрытиям [5-10] была установлена взаимосвязь ряда особенностей наблюдаемых перекрытий с некоторыми свойствами структуры генетического кода. Для получения основного результата по такой взаимосвязи проводилось математическое моделирование полного множества указанных перекрытий. Показано [11], что для наиболее протяженного генетического перекрытия, обнаруженного в вирусе GSHV и содержащего 428 кодонов [14], потенциально возможное число перекрытий составляет $\sim 10^{746}$. В ходе исследования были выделены все возможные наборы аминокислот, для которых имеет место перекрываемость [11,12]. Под перекрываемостью будем понимать такие случаи, когда в гене, соответствующем сдвинутому состоянию, не может возникнуть ни один из кодонов *ter*. Назовем перекрываемостью полной, если она имеет место для последовательности любой протяженности с произвольным сочетанием аминокислот. Математическое моделирование полных множеств перекрытий для случаев 3-5 (см. рис.3), которое проводилось по аналогии со случаями 1-2 (см. [2,3,11,12]), позволило установить неизвестное ранее свойство K_0 , которое сформулируем в виде теоремы.

Теорема 2. *Стандартный генетический код допускает полную перекрываемость для каждого из случаев 1-5 за исключением последовательностей, содержащих хотя бы одну из пар аминокислот в следующих трех случаях перекрытия: в случае 2 это 5 пар:*

$$\text{MetMet, MetAsn, MetLys, MetIle, MetThr,} \quad (10)$$

в случае 3 это 6 пар:

$$\text{PheTyr, TyrTyr, HisTyr, AsnTyr, AspTyr, CysTyr,} \quad (11)$$

в случае 5 это 5 пар:

$$\text{PheMet, PheAsn, PheLys, PheIle, PheThr.} \quad (12)$$

Тем самым на основе теоремы 2 получена интегральная характеристика, которую обозначим через p и которая для стандартного кода равна 16. Для кодов, отклоненных от кода стандартного, такая характеристика может измениться (см. ниже).

Пример использования теоремы 2 приведем для последовательности из 4-х аминокислот: PheMetAsnTyr, которая с учетом всех кодировок Phe, Met, Asn, Tyr порождает 8 генов. Однако сдвиг любого из таких генов всегда дает запреты для 3-х случаев: для случая 1 из-за пары MetAsn из (10), так как при этом возникает кодон TGA, для случая 3 из-за пары AsnTyr из (11), так как в этом случае возникает кодон TAG либо TAA, и для случая 5 из-за пары PheMet из (12); в этом случае возникает кодон TGA либо TAA.

Следствие 1. Код K_0 допускает полную перекрываемость для каждого из случаев 1-5, если перекрываемые последовательности не содержат какую-либо из 16 пар аминокислот.

кислот из наборов (10)-(12).

Следствие 2. Полная перекрываемость для каждого из случаев 1-5 имеет место для любого кода у которого 63 кодона являются смысловыми, из которых 61-кодон K_0 , а один кодон – TAG – терминаторным.

Анализ полученных решений позволил сделать два утверждения.

Таблица 1. Дополнительные запреты на генетические перекрытия в случаях 1-5 (указано в скобках, см. рис.3), которые возникают для гипотетических кодов K_1 - K_4 , каждый из которых образован перестановкой в K_0 только одного смыслового кодона в семейство терминаторных кодонов (ter).

K_1	K_2	K_3	K_4
GGC (Gly) → ter	CCA (Pro) → ter	GCC (Ala) → ter	ATA (Ile) → ter
MetAla (1) AT GGCN ter	Trp (4) TGG ACC Ter	MetPro (1) AT GGCN Ter	IleTyr (3) ATYTA TAXAT ter
TrpAla (1) T GGGCN ter		TrpPro (1) T GGCCN Ter	IleMet (5) ATYATG TAXTAC ter
TrpHis (2) T GGCAY ter		TrpAla (3) T GGGCN ACCCG Ter	IleAsn (5) ATYAA Y TAXTT ter
TrpGln (2) T GGCAX ter		MetAla (3) AT GGCN TACCG Ter	IleLys (5) ATYAA X TAXTT ter
MetPro (3) AT GGCN TACGG ter			IleIle (5) ATYATY TAXTA ter
TrpPro (3) T GGCCN ACCCG ter			IleThr (5) ATYACN TAXTG ter

1. Рассмотрим вопрос о том, как изменится перекрываемость при отклонениях кода от K_0 . Постановка такого вопроса была связана с известным положением «Код, по-видимому, был «выбран» произвольно...» [15, с.18]. Наше исследование этого положения не

подтверждает. Для обоснования нашей позиции были изучены вопросы перекрываемости для ряда гипотетических кодов, отклоненных от K_0 . В табл.1 приведены данные о дополнительных запретах на перекрываемость, возникающих всего лишь при перестановке одного кодона в K_0 . Каждая из таких перестановок соответствует расширению набора кодонов *ter* на 1 и такой набор становится четырехзначным. Следует отметить, что такие размерности *ter* наблюдаются в некоторых девиантных кодах [16]. Из табл.1 следует, что для K_1 ($GGC(Gly) \rightarrow ter$) возникают запреты при использовании последовательностей, содержащих 7 пар аминокислот: MetAla, TrpAla, TrpHis, TrpGln, TrpPro, MetPro, TrpPro. Для каждой из пары аминокислот этого набора возникает кодон *ter*(GGC) для разных случаев перекрываемости: для случаев 1,2,3, которые указываются в скобках в первом столбце табл.1. То есть возникают запреты в том числе и в случае 1, для которого согласно теореме имеет место полная перекрываемость для K_0 . Подобные случаи были установлены ранее для перестановок CGA(Arg) $\rightarrow ter$ и CAA(Gln) $\rightarrow ter$, для каждой из которых (см. табл.1 из [11]) возникали несовпадающие запретные наборы, содержащие одинаковое количество пар аминокислот – 12. Это число соответствует случаю 1 (см. рис.3); оно сравнимо по величине с полным числом запретов на перекрывание (16), которое имеет место для всех пяти случаев из рис.3, которые, согласно теореме, соответствуют K_0 . Вновь обратимся к табл.1. Одиночные перестановки, представленные в ней, приводят в целом к дополнительным запретам по всем пяти случаям перекрываемости. Для K_2 (CCA(Pro) $\rightarrow ter$) такой запрет возникает в случае 4, для K_3 (GCC(Ala) $\rightarrow ter$) – в случаях 1 и 3, для K_4 (ATA(He) $\rightarrow ter$) – в случаях 3, 5. Итак, одиночные перестановки в K_0 (полное их число оказалось более 100) приводят к запретам на перекрываемость, которых нет для K_0 . Тем самым можно утверждать, что K_0 «выбран» произвольно. Более того, можно утверждать, что одним из решающих факторов в «выборе» структуры K_0 явилась его способность к почти полной перекрываемости пар генов для каждого из случаев 1-5, т.е. для всех случаев, допускаемых структурой ДНК. Подчеркнем, что именно «выбором» такой, на первый взгляд необычной вырожденности K_0 , мы и обязаны указанному свойству почти полной перекрываемости. Некоторые из элементов такой вырожденности были изучены ранее для перекрытий в случае 1: это особенности семейств Ser, Leu, Arg (см.[8], а также [6,7,10]), необычность семейства *ter* [9]. Однако для K_0 почему-то нет полной перекрываемости, хотя код, близкий к K_0 , таким свойством обладает (см. следствие 2). Этот вопрос требует отдельной проработки с учетом других функций кодонных семейств K_0 , помимо перекрываемости генов.

2. Нами был также поставлен вопрос о возможной взаимосвязи ограничений на перекрываемость (10)-(12) с вариабельностью кода, наблюдаемой у ряда организмов. Анализ показал, что такая взаимосвязь существует и она выражается в том, что для ряда девиантных кодов (примеры для некоторых из них, обнаруженных в митохондриальных ДНК, представлены на рис.4) перестановки кодонов приводят к возможности построения генетических перекрытий, запретных для K_0 и установленных теоремой 2. В каждом из четырех фрагментов перекрытий, приведенных на рис.4 (построен на основе экспериментальных данных из [17-20]), указывается роль одной и той же перестановки: TGA(*ter*) \rightarrow Trp. Оказалось, что такая перестановка делает возможными перекрытия для пар MetAsn (рис.4.1, этот случай соответствует ДНК митохондрии человека), MetMet (рис.4.2), MetThr (рис.4.3) и MetLys (рис.4.4), указанных в формулировке теоремы 2 (см.

(10)) как запретные на перекрываемость для K_0 . Отметим, что по современным представлениям происхождение перекрывающихся генов может быть связано с процессом эволюции молекул ДНК, когда в определенных ситуациях могли сказываться ограничения на физический размер генома. Данное исследование показывает, что такие ограничения могут носить приоритетный характер по сравнению с постоянством генетического кода; в рассматриваемых случаях рис.4 вариации кода приводят к возможности построения запретных для K_0 генетических перекрытий и как следствие к сокращению физического размера геномов. Тем самым не подтверждается положение из [15] о том, что «в генетическом коде митохондрий могут происходить случайные перемены».

1 (human)

АТФаза6→ **MetAsn...**

URF A6L→...**Trp...**
...**ATGAA...**

8528

2 (*Drosophila yakuba*)

АТФаза6→ **MetMet...**

URF A6L→...**Trp...**
...**ATGATG...**

4067

3 (*Paracentrotus lividus*)

АТФаза6→ **MetThrMetThr...**

АТФаза8→...**TrpGlnTrp...**
...**ATGACAATGAC...**

8680

4 (*Apis mellifera ligustica*)

АТФаза6→ **MetLys...**

АТФаза8→...**Trp...**
...**ATGAA...**

4585

Рис.4. Фрагменты генетических перекрытий, обнаруженные в митохондриях четырех организмов. Это перекрытия в одной цепи ДНК, в каждом из которых оказалась кодировка, отличная от K_0 . Названия белков даны по публикациям [17-20] соответственно, а число под нуклеотидом указывает его номер в геноме. Устанавливается влияние одной и той же вариации кода TGA(ter)→Trp, которая имеет место во всех четырех девиантных кодах. Другие вариации таких кодов не оказывают влияния на представленные фрагменты перекрытий и в данной работе не рассматриваются (см. [8]). Оказалось, что вариация делает возможными генетические перекрытия, запретные для K_0 . Это имеет место при использовании четырех (выделено жирным шрифтом) пар аминокислот: 1 – MetAsn, 2 – MetMet, дважды MetThr в 3 и 4 – MetLys, которые указываются в (1) как запретные для K_0 . Согласно (10) таких пар всего 5 для перекрытий из одной цепи ДНК. См. текст.

Изучался также вопрос об изменении характеристики p , введенной на основе теоремы 2, для некоторых кодов, отклоненных от стандартного кода. Было показано [21], что для трех таких кодов, соответствующих перекрытиям на рис.2, обнаруживается уменьшение значения p в 2.3-3.2 раза. Именно подобное уменьшение и приводит к возможности построения перекрытий генов, запрещенных при использовании стандартного кода.

Многостороннее изучение особых способов записи генетической информации, всех типов перекрывающихся генов, когда один и тот же участок ДНК кодирует от двух до шести белков (вместо традиционного одного), вступает в новый период. После первых сообщений о значительном числе подобных генов в самой большой из хромосом человека [22] появились обработанные данные по всему человеческому и другим большим геномам. По данным [23] в человеческом геноме, который был расшифрован относительно недавно, было обнаружено около 1700 перекрытий пар генов. Полученные результаты по человеческому и другим большим геномам вызвали удивление у специалистов, т.к. устоявшимся мнением было то, что перекрытия генов являются свойством только малых геномов. У таких геномов кодирующая часть (область ДНК, кодирующая белковые последовательности) может достигать 100% (см рис.1 из [2]), в то время как в человеческом геноме – не более 3%. Для геномов малых размеров, за счет генетических перекрытий, размер генома может сократиться в полтора раза [2], а для больших геномов такое сокращение просто ничтожно. Тем самым возникают новые вопросы, требующие своего решения.

3. Интегральные характеристики генетического кода

Как известно, кодирование одной и той же аминокислоты может осуществляться несколькими триплетами. Это свойство было названо вырожденностью генетического кода. Встает вопрос: как же природа использует эту вырожденность? Нами выше раздельно изучались две функции вырожденности кода: блокировка пяти последовательностей, альтернативных гену, который не перекрывается другим геном (п.1) и перекрываемость пар генов (п.2). Были установлены интегральные характеристики генетического кода, которые использует природа для выполнения названных функций. Ниже устанавливается связь между указанными характеристиками. Показано, что малость одной интегральной характеристики, обнаруженной для перекрывающихся генов, является следствием более общего принципа, который использует природа для «записи» не перекрывающихся генов. Тем самым устанавливается один из критериев, который мог быть использован для «выбора» окончательной структуры генетического кода.

На основе теоремы из п.1, а также с учетом последующего анализа, была введена в рассмотрение двухкомпонентная интегральная характеристика генетического кода q из (9), которая устанавливает количество и тип пар аминокислот, могущих вызывать названные блокировки. Характеристика q состоит из двух компонент. Первая компонента – $q_{min}=31$ – соответствует так называемым неизбежным блокировкам, а вторая – $q_{max}=210$ – максимально возможному числу блокирующих пар. На рис.5 приводится полный перечень блокирующих пар, полученных на основе п.1. Пары, соответствующие неизбежным блокировкам, указаны буквой i от слова inevitable – неизбежный, а оставшиеся 179 пар аминокислот указаны буквой v от слова virtual – возможный. Тем самым введе-

ны в рассмотрение два набора пар: *i*-пары и *v*-пары. Обратимся к полному набору *i*- пар, который представлен в табл.2. В столбце 2 этой таблицы указываются номера РС для которых (потенциально) может иметь место блокировка. Число таких РС в зависимости от пары меняется от 1 до 4. Разделим множество этих пар на два подмножества. К первому отнесем *i*-пары, неизбежно блокирующие одну и ту же РС. Таких пар оказалось всего 16 – это первые 16 пар из табл.2. Имеем РС2 неизбежно блокируют 5 пар аминокислот, которые совпадают с набором (10), определенным в теореме 2, поскольку в РС2 образуется кодон *ter* –TGA. Для РС3 имеем 6 блокирующих пар, совпадающих с набором (11) согласно той же теореме, в РС3 образуется один из кодонов *ter*: TAA или TAG. Особо следует сказать о блокировке для пяти пар с номерами 12-16, которые совпадают с парами из (12), определенными также в теореме 2. Отметим лишь, что каждая из этих последних пар неизбежно блокирует РС5, поскольку в РС5 образуется один из кодонов *ter*: TAA или TGA. Однако в парах PheAsn, PheLys помимо РС5 в табл.2 указывается также РС1. Однако последняя РС не соответствует неизбежной блокировке в отличие от РС5. Таким образом, пары 1-16 из табл.2 образуют набор пар аминокислот, запретных для перекрытий двух генов и были установлены в п.2. На основе п.2 была введена в рассмотрение числовая характеристика, которая была обозначена буквой *p* и которая соответствует числу различающихся блокировочных пар из (10)-(12), имеем $p=16$. Тем самым использование подмножества неизбежных блокировок позволяет установить связь двух теорем: 1 и 2. Таким образом, имеет место неравенство:

$$0 \leq p \leq q_{min}. \quad (13)$$

Иными словами интегральная характеристика генетического кода *p* не является независимой, а определяется выбором характеристики q_{min} , которая используется в решении совершенно другой задачи – в блокировке неперекрывающихся генов.

Таблица 2. Полный перечень *i*-пар (столбец 1) с указанием номеров РС, которые блокируются (столбец 2).

№	1	2	№	1	2
1	MetMet	2	17	IleMet	2,5
2	MetAsn	2	18	ValMet	2,5
3	MetLys	2	19	LeuMet	2,4,5
4	MetIle	2	20	IleAsn	1,2,5
5	MetThr	2	21	ValAsn	1,2,5
6	PheTyr	3	22	LeuAsn	1,2,4,5
7	TyrTyr	3	23	IleLys	1,2,5
8	HisTyr	3	24	ValLys	1,2,5
9	AsnTyr	3	25	LeuLys	1,2,4,5
10	AspTyr	3	26	IleIle	2,5
11	CysTyr	3	27	ValIle	2,5
12	PheMet	5	28	LeuIle	2,4,5
13	PheAsn	1,5	29	IleThr	2,5
14	PheLys	1,5	30	ValThr	2,5
15	PheIle	5	31	LeuThr	2,4,5
16	PheThr	5			

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Σ	
1 Met	i		i									i	i	v	v	v		v	i		9	
2 Trp																		v	v		2	
3 Phe																		v	v		2	
4 Tyr			i	i	i	i	i	i				v	v	v	v	v	v	v	v	v	v	15
5 His			v	v	v	v	v	v				v	v	v	v	v	v	v	v	v	v	15
6 Asn	i		i	v	v	v	v	v				i	i	v	v	v	v	v	i	v	16	
7 Asn			v	v	v	v	v	v				v	v	v	v	v	v	v	v	v	v	15
8 Cys																		v	v		2	
9 Gln			v	v	v	v	v	v				v	v	v	v	v	v	v	v	v	v	15
10 Lys	i		i	v	v	v	v	v				i	i	v	v	v	v	v	i	v	16	
11 Glu			v	v	v	v	v	v				v	v	v	v	v	v	v	v	v	v	15
12 Ile	i		i									i	i	v	v	v		v	i		9	
13 Val												v	v					v	v		4	
14 Pro																		v	v		2	
15 Thr	i		i									i	i	v	v	v		v	i		9	
16 Ala												v	v					v	v		4	
17 Gly												v	v					v	v		4	
18 Ser	v	v	v	v	v	v	v	v	v	v	v	v	v	v	v	v	v	v	v	v	v	20
19 Leu	v	v	v	v	v	v	v	v	v	v	v	v	v	v	v	v	v	v	v	v	v	20
20 Arg	v		v	v	v	v	v	v				v	v	v	v	v	v	v	v	v	v	16
Σ	8	2	13	10	10	10	10	10	2	2	2	16	16	13	13	13	10	20	20	10	210	

Рис.5. Упрощенный рисунок, указывающий 210 пар аминокислот из 400 возможных, потенциально блокирующих PC1-PC5 (построен на основе теоремы 1 из п.1). По оси абсцисс указывается первая аминокислота из пары, по оси ординат – вторая. Символ 'i' соответствует паре аминокислот, для которой блокировка неизбежна, то есть возникает при любых кодировках (от inevitable – неизбежный). Символ 'v' соответствует паре аминокислот, для которых блокировка возникает не при всех кодировках (от virtual – виртуальный). Число i-пар=31, v-пар=179. Речь идет о потенциально возможных блокировках всего одной PC (таких пар 158) либо более чем одной (таких пар 52). Оказалось, что для стандартного кода не существует пар аминокислот, потенциально блокирующих все 5 PC: PC1-PC5. Обнаружены всего 4 пары аминокислот (LeuAsn, LeuLys – это i-пары; LeuSer, LeuArg – это v-пары), потенциально блокирующие 4 PC: PC1, PC2, PC4, PC5, но при этом никакая из указанных 52 пар не может одновременно давать блокировку более двух PC. Число аминокислот, участвующих в блокирующих парах в качестве первой компоненты, указано в нижней строке (минимум -2 для Trp, Gln, Lys, Glu, максимум -20 для Ser, Leu), а в качестве второй компоненты – в правом столбце (минимум -2 для Trp, Phe, Cys, Pro, максимум -20 для Ser, Leu). Среднее значение указанного числа около 10.

При рассмотрении только одной задачи – перекрытия пар генов – в п.2 был сделан вывод о том, что генетический код был «выбран» под перекрытия генов, поскольку только 16 пар из 400 возможных запретны для перекрытий. Это справедливо для всех 5 способов парных перекрытий генов, разрешенных структурой ДНК. Однако при рассмотрении двух задач – блокировки генов и перекрытия генов и двух интегральных характеристик p и q , оказалось, с учетом связи (13), что генетический код был ориентирован на «выбор» двухкомпонентной интегральной характеристики (9), одна из компонент которой согласно неравенству (13) определяет область значения другой интегральной

характеристики p . Таким образом, малость интегральной характеристики p является следствием более общего принципа, связанного с выбором всего набора неизбежных пар, создающих блокировки. Т.е. пары аминокислот, соответствующие характеристике p , могут быть «выбраны» только из этого ограниченного набора, соответствующего q_{min} , а не из полного набора 400 пар аминокислот. По какому критерию произошел такой «выбор», пока остается неясным.

Отметим, что впервые рассмотренные здесь вопросы были кратко представлены в [14,24, 25].

Вопросы применения изложенной теории будут даны в последующих публикациях. Кратко это применение было представлено в [26], где рассматривались результаты расчетов для более 200 000 генов, отвечающих 12 большим геномам, в том числе генома человека, содержащего 25 613 структурных генов. Другая принципиально важная задача была связана с рассмотрением интегральных и ряда других характеристик для всех генов в живой клетке [27], где, как известно, гены записаны двумя различающимися генетическими кодами.

СПИСОК ЛИТЕРАТУРЫ

1. *Козлов Н.Н.* Математический анализ особого способа записи генетической информации // Математическое моделирование, 1995, т.7, №12, с.33-47.
2. *Козлов Н.Н.* Математический анализ перекрывающихся генов и структура генетического кода // Математическое моделирование, 2000, т.12, №7, с.97-101.
3. *Козлов Н.Н.* Один способ хранения генетической информации // Математическое моделирование, 2002, т.14, №8, с.72-78.
4. *Льюин Б.* Гены. – М.: Мир, 1987, 544 с.
5. *Козлов Н.Н., Кугушев Е.И.* Изучение модели сворачивания тРНК во вторичную структуру // Математическое моделирование, 1993, т.5, №6, с.24-55.
6. *Козлов Н.Н.* Об особом способе записи генетической информации // ДАН, 1994, т.337, №1, с.158-161.
7. *Козлов Н.Н.* Молчащие мутации в области перекрывания генов // ДАН, 1996, т.350, №5, с.699-703.
8. *Козлов Н.Н.* Перекрывающиеся гены и генетический код // ДАН, 1997, т.355, №6, с.830-833.
9. *Козлов Н.Н.* Терминаторные кодоны в генетических перекрытиях // ДАН, 1998, т.360, №4, с.550-553.
10. *Козлов Н.Н.* О востребованности каждого из 64 кодонов в генетических перекрытиях // ДАН, 1999, т.367, №4, с.544-547.
11. *Козлов Н.Н.* К вопросу о произвольности «выбора» генетического кода // ДАН, 1999, т.369, №4, с.553-556.
12. *Козлов Н.Н.* Анализ полного множества перекрывающихся генов // ДАН, 2000, т.373, №1, с.108-111.
13. *Козлов Н.Н.* Перекрывающиеся гены и вариабельность генетического кода // ДАН, 2000, т.375, №6, с.824-827.
14. *Козлов Н.Н.* Теорема для генетического кода // ДАН, 2002, т.382, №5, с.593-597.
15. *Албертс Б., Брей Д., Льюис Дж., Рэфф М., Робертс К., Уотсон Дж.* Молекулярная биология клетки. Т.1. – М.: Мир, 1994, 514 с.
16. *Jukes T.H.* // *Experientia*, 1990, v.46, №11/12, p.1149-1157.

3 Математическое моделирование, №9

17. *Anderson S., Bankier A.T., Barrell B.G. et al.* // *Nature*, 1981, v.290, p.457-464.
18. *Clary Douglas O., Wolstenholme David R.* // *J. Mol. Evolut.*, 1985, v.22, p.252-271.
19. *Cantatore P., Roberti M., Rainaldi G. et al.* // *J. of Biological Chemistry*, 1989, v.264, №19, p.10965-10975.
20. *Crozier R., Crozier Y.* // *Genetics Society of America*, 1993, v.133, p.97-117.
21. *Козлов Н.Н.* Применение теоремы для генетического кода // *ДАН*, 2004, т.396, №6, с.740-745.
22. *Gregory S.G., Barlow K.F., McLay K.E., et al.* // *Nature*, 2006, v.441, p.315-321.
23. *Nakayama T., Asai S., Takahashi Y., et al.* // *NJBS*, 2007, v.3, №1, p.14-19.
24. *Козлов Н.Н.* Математический анализ структурных генов // *ДАН*, 2007, т.412, №5, с.610-613.
25. *Козлов Н.Н.* Интегральные характеристики генетического кода // *ДАН*, 2007, т.417, №1, с.30-33.
26. *Козлов Н.Н., Грязнов С.С.* Некоторые расчетные характеристики больших геномов // *ДАН*, 2007, т.417, №6, с.732-737.
27. *Козлов Н.Н.* Математический анализ девиантности генетического кода // *ДАН*, 2007, т.415, №4, с.441-445.

Поступила в редакцию 26.11.09