



ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 26 за 2010 г.



Козлов Н.Н.

Расчет генетического кода

Рекомендуемая форма библиографической ссылки: Козлов Н.Н. Расчет генетического кода
// Препринты ИПМ им. М.В.Келдыша. 2010. № 26. 23 с. URL:
<http://library.keldysh.ru/preprint.asp?id=2010-26>

ISSN 2071-2898



**Ордена Ленина
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В. Келдыша
Российской академии наук**

Н.Н.Козлов

РАСЧЕТ ГЕНЕТИЧЕСКОГО КОДА

Препринт № 26

Москва

Н.Н. Козлов. Расчет генетического кода. Препринт ИПМ им. М.В. Келдыша РАН, М., 2010, 23 с.

Математический анализ генетических расшифровок проводился автором в течение ряда лет. Причем одним из главных объектов исследования были перекрывающиеся гены, открытые в 1976 году. Это случаи, когда один и тот же участок ДНК кодирует две или более белковых последовательностей. Многочисленные случаи выявленных генетических перекрытий, позволили поставить ряд математических задач, которые были успешно решены. Подробное изложение этих задач приводится в авторской монографии. «Математический анализ генетического кода», (М., БИНОМ, 2010, 223 с.) Такие задачи дали возможность достаточно глубоко проникнуть в структуру генетического кода и ее взаимосвязи с перекрывающимися генами. В итоге была поставлена новая задача: расчет генетического кода на основе аминокислотных последовательностей, записывающих перекрывающиеся гены. Об одном подходе к решению этой задачи говорится в данной работе.

N.N. Kozlov. Computation of the genetic code. Preprint KIAM RAS, Moscow, 2010, 23p.

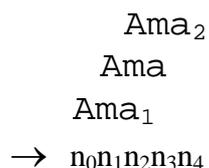
The mathematical analysis of genetic decodings was spent by the author for a number of years. And the overlapping genes opened in 1976 were one of the main objects of research. These are cases when the same site of DNA codes two or more protein sequences. Numerous cases of the revealed genetic overlappings, permitted to put a number of mathematical problems which have been successfully solved. The detailed statement of these problems is resulted in the author's monography. «The mathematical analysis of a genetic code», (M, BINOM, 2010, 223 pp.) Such problems have given the chance to get into structure of a genetic code and its interrelation with overlapping genes deeply enough. As a result the new problem has been put: computation of a genetic code on a basis of amino acids sequences which are writing down overlapping genes. About one approach to the decision of this problem it is told in the given work.

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 10-01-00145, 08-01-00042), а также гранта ведущих научных школ (НШ – 6700. 2010.1).

Как известно, генетический код был окончательно установлен экспериментально в 1966 году. Сразу после этого началась расшифровка генов с использованием генетического кода. Главный объект исследований автора также гены, но перекрывающиеся, открытые в 1976 году и обнаруживаемые в настоящее время в значительных количествах в больших геномах. Перекрывающиеся гены это случаи, когда один и тот же участок ДНК кодирует более одной (вплоть до 6-и) белковой последовательности. В ходе исследований были решены ряд задач для области ДНК с перекрытиями генов: степень жесткости связи белковых последовательностей и этой области [1,2], роль нерегулярностей кода [3], востребованность каждого из смысловых кодонов [4], способ построения любого перекрытия пар генов из одной цепи ДНК [5,6], уменьшение размера генома, за счет отклонений кода от его стандартной структуры [7], математический анализ всех 5-и способов парных перекрытий генов для стандартного [8], а также всех 14-и девиантных кодов [9], введение интегральных характеристик кода как для перекрывающихся, так и для не перекрывающихся генов и установление их взаимосвязи [10-13], установление математической аналогии генетических перекрытий и стеблей вторичной структуры матричных РНК [12], теория элементарных генетических перекрытий [14], а также роль неоднозначностей во множествах таких перекрытий [15]. Подробное изложение этих задач в их взаимосвязи представлено в авторской монографии [16]. В ходе решения указанных задач возникла мысль попытаться расчетным путем установить структуру генетического кода, или найти такой тип математических преобразований, на основе которых можно было бы вычислить генетический код. Об одном подходе к решению этой задачи говорится в данной работе.

Ранее нами было введено понятие элементарного перекрытия применительно к перекрытиям пар генов [5,6]. Обобщим это понятие для трех генов, принадлежащих одной и той же цепи ДНК. Под термином элементарное перекрытие понимаем, перекрытие для кодонов одиночных аминокислот по максимальному числу позиций. На рис.1А для аминокислоты Ама,

(А)



(Б)

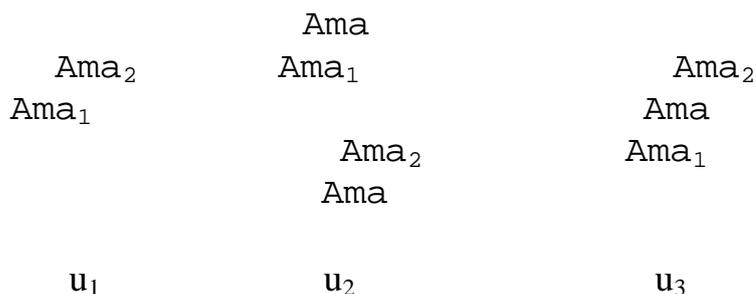


Рис.1. А. Для аминокислоты Ама, кодируемой триплетом $p_1p_2p_3$, имеют место 2 альтернативных аминокислоты Ама_1 и Ама_2 . Кодировки их образованы сдвигами -1 , $+1$ в той же цепи ДНК (\rightarrow). Кодон $p_0p_1p_2$ для Ама_1 перекрывается с кодоном $p_1p_2p_3$ для Ама – перекрытие содержит два нуклеотида p_1p_2 ; кодон $p_2p_3p_4$ для Ама_2 перекрывается с кодоном $p_1p_2p_3$ для Ама – перекрытие содержит два нуклеотида p_2p_3 . В итоге тройное перекрытие содержит всего одну общую позицию p_2 . Каждая из приведенных трех аминокислот принадлежит тем наборам A^0 из (1), кодировки которых допускают такие перекрытия.

Б. Элементы множеств сочетаний аминокислот, образованные на основе элементарного перекрытия из рис1А. Слева – один элемент множества U_1 , в центре – два элемента множества U_2 и справа – один элемент множества U_3 .

кодируемой триплетом $p_1p_2p_3$ указаны альтернативные аминокислоты A_{α_1} и A_{α_2} , кодировки которых $p_0p_1p_2$ и $p_2p_3p_4$ соответственно, образованы сдвигами на -1 и $+1$ нуклеотид в той же цепи ДНК (\rightarrow). Предполагается при этом, что все значения $p_0 - p_4$ принадлежат каноническому набору из четырех нуклеотидов. На основе рис.1А можно построить три вида сочетаний аминокислот, представленных на рис.1Б и обозначенных соответственно u_1 , u_2 , u_3 : одно u_1 для перекрытия по одной позиции, два u_2 для перекрытия по двум позициям, и одно u_3 для перекрытия 3-х аминокислот. Следует отметить, что в генетических экспериментах более 10 лет назад нами были обнаружены все допустимые генетическим кодом элементы u_1 , u_2 , u_3 , которые будем рассматривать соответственно как элементы множеств U_1 , U_2 , U_3 . Число элементов U_2 равно 160, а элементов $U_3 - 307$; полный их перечень представлен в Приложении стр. 16-23. Что касается множества U_1 , то в ходе решения были использованы лишь небольшая часть элементов этого множества и его обобщенные характеристики, которые приводятся в тексте.

ПОСТАНОВКА ЗАДАЧИ. Пусть имеем набор из 4 букв: $N: a, b, c, d$, а также триплеты – любые тройки из этих букв, всего их 64. При этом каждая из 20 канонических аминокислот может кодироваться произвольным сочетанием таких триплетов. Задача состоит в поиске всех генетических кодов, отвечающих всем элементам, обозначенных выше трех множеств U_1 , U_2 , U_3 , соответствующих генетическим экспериментам.

РЕШЕНИЕ ЗАДАЧИ. Используем стандартные трехбуквенные сокращения для каждой из 20 аминокислот, указанные в табл.1. Имеем набор A^0

$$A^0: A_{\alpha_i}, \quad i \in (1, 20). \quad (1)$$

Введем определение. Назовем сочетание аминокислот, построенное на основе элементарного генетического перекрытия однородным, если в нем участвует одна и та же аминокислота. Для однородных u_3 имеет место

Таблица 1. Некоторые данные множеств U_2 и U_1 для аминокислот $Ama_i, i \in (1,20)$.

i	Ama_i	U_2		U_1		K
		m_{12}	m_{23}	m_{1-3}	m_{3-1}	
1	Met	4(Tyr,His,Asn,Asp)	2(Trp,Cys)	3(Lys,Pro,Gly)	1(Gly)	abd
2	Trp	3(Met,Val,Leu)	1(Gly)	3(Phe,Pro,Gly)	1(Gly)	bdd
3	Phe	4(Phe,Ile,Val,Leu)	3(Phe,Ser,Leu)	3(Phe,Pro,Gly)	2(Phe,Pro)	bbY
4	Tyr	3(Ile,Val,Leu)	3(Met,Ile,Thr)	3(Phe,Pro,Gly)	2(Phe,Pro)	baY
5	His	4(Pro,Thr,Ala,Ser)	3(Met,Ile,Thr)	3(Phe,Pro,Gly)	2(Phe,Pro)	caY
6	Asn	3(Gln,Lys,Glu)	3(Met,Ile,Thr)	3(Lys,Pro,Gly)	2(Phe,Pro)	aaY
7	Asp	2(Gly,Arg)	3(Met,Ile,Thr)	3(Lys,Pro,Gly)	2(Phe,Pro)	daY
8	Cys	3(Met,Val,Leu)	2(Val,Ala)	3(Phe,Pro,Gly)	2(Phe,Pro)	bdY
9	Gln	4(Pro,Thr,Ala,Ser)	4(Asn,Lys,Ser,Arg)	3(Phe,Pro,Gly)	2(Lys,Gly)	caX
10	Lys	3(Gln,Lys,Glu)	4(Asn,Lys,Ser,Arg)	3(Lys,Pro,Gly)	2(Lys,Gly)	aaX
11	Glu	2(Gly,Arg)	4(Asn,Lys,Ser,Arg)	3(Lys,Pro,Gly)	2(Lys,Gly)	daX
12	Ile	4(Tyr,His,Asn,Asp)	4(Phe,Tyr,Ser,Leu)	3(Lys,Pro,Gly)	3(Phe,Lys,Pro)	abM
13	Val	4(Cys,Gly,Ser,Arg)	6(Trp,Phe,Tyr,Cys,Ser,Leu)	3(Lys,Pro,Gly)	4	dbN
14	Pro	4(Pro,Thr,Ala,Ser)	5(His,Gln,Pro,Leu,Arg)	3(Phe,Pro,Gly)	4	ccN
15	Thr	4(Tyr,His,Asn,Asp)	5(His,Gln,Pro,Leu,Arg)	3(Lys,Pro,Gly)	4	acN
16	Ala	4(Cys,Gly,Ser,Arg)	5(His,Gln,Pro,Leu,Arg)	3(Lys,Pro,Gly)	4	dcN
17	Gly	3(Trp,Gly,Arg)	5(Asp,Glu,Val,Ala,Gly)	3(Lys,Pro,Gly)	4	ddN
18	Ser	7(Phe,Gln,Lys,Glu,Ile,Val,Leu)	7(His,Gln,Val,Pro,Ala,Leu,Arg)	4	4	bcN adY
19	Leu	8(Phe,Ile,Val,Pro,Thr,Ala,Ser,Leu)	6(Trp,Phe,Tyr,Cys,Ser,Leu)	3(Phe,Pro,Gly)	4	cbN bbX
20	Arg	7(Gln,Lys,Glu,Pro,Thr,Ala,Ser)	5(Asp,Glu,Val,Ala,Gly)	4	4	cdN adX

Обозначения. Первые 4 столбца указывают на число перекрытий, а в скобках перечень аминокислот для перекрытий: по 1 и 2 основаниям (m_{12}), по 2 и 3 основаниям (m_{23}) по 1 основанию с третьим основанием (m_{1-3}), по 3 основанию с первым основанием (m_{3-1}). В столбцах 3 и 4 речь идет только о перекрытиях с Lys, Phe, Pro, Gly, поэтому это число не может быть более 4-х. В последнем столбце приводятся вычисленные значения K. Обозначения: X: a,d; Y: в, с; M: а, в, с; N: а, в, с, d.

Свойство. Пусть кодировки Ama для однородного u_3 имеют одно из трех следующих представлений:

$$n_1N_1N_2, \quad N_3n_2N_4, \quad N_5N_6n_3, \quad (2)$$

где малыми буквами обозначены единичные составляющие набора N, а большими – какие-либо подмножества этого набора, вплоть до N. Тогда однородный u_3 может существовать только при использовании хотя бы одного базового триплета или триплета с тремя одинаковыми буквами.

Для доказательства осуществим последовательно подстановки каждого из представлений (2) в u_3 : "

	Ama	Ama	Ama	
	Ama	Ama	Ama	(3)
	Ama	Ama	Ama	
Н . П .	$n_1n_1n_1n'_1n'_2$	$n'_3n_2n_2n_2n'_4$	$n'_5n'_6n_3n_3n_3$	

где n'_i - одиночная составляющая набора N_i , где $i \in (1,6)$, а строка н.п. - нуклеотидные последовательности, которые образуются после указанной подстановки. В первом случае в (3) базовый кодон $n_1n_1n_1$ был использован при кодировке аминокислоты Ama из нижней позиции, во втором - $n_2n_2n_2$ - из средней позиции, а в третьем - $n_3n_3n_3$ - из верхней позиции.

Обратимся к однородным u_3 из множества U_3 , которых оказалось 4:

Lys	Phe	Pro	Gly	
Lys	Phe	Pro	Gly	(4)
Lys	Phe	Pro	Gly	

В рамках предположения, указанного в Свойстве предлагается следующий пошаговый процесс поиска генетического кода; см. Таблицу 2.

Шаг 1. Аминокислотам из (4) присвоим соответствующие базовые кодоны. Такое присвоение не является однозначным. Однако в нашем подходе набор букв N: a, b, c, d не сопоставлен с каноническим набором 4-х нуклеотидов; об этом речь пойдет в конце работы. Поэтому далее будем оперировать лишь

с одним из представлений для аминокислот из (4), которым присвоим соответственно следующие базовые триплеты:

$$\text{Lys: aaa, Phe: bbb, Pro: ccc, Gly: ddd} \quad (5)$$

Для дальнейших расчетов обратимся к некоторым обобщенным данным по множествам U_2 и U_1 , которые приводятся в таблице 1.

Шаг 2. Из табл.1 следует, что, как в столбце m_{12} (указывается число и дается перечень перекрываемых аминокислот по 1и2 основаниям), так и в столбце m_{23} (аналогичные данные по 2 и 3 основаниям) не содержатся взаимных перекрытий между аминокислотами из (4). Такие перекрытия имеют место только по одной позиции и принадлежат множеству U_1 . Имеем:

$$\begin{array}{cccc}
 \text{Lys} & \text{Phe} & \text{Pro} & \text{Gly} \\
 \text{Lys} & \text{Phe} & \text{Lys} & \text{Lys} \\
 (\text{Gly}) & (\text{Pro}) & (\text{Pro}) & (\text{Phe}) \\
 & & (\text{Gly}) & (\text{Pro}) \\
 & & (\text{Phe}) & (\text{Gly})
 \end{array} \quad (6)$$

$$\begin{array}{cccc}
 \text{Lys} & \text{Phe} & \text{Pro} & \text{Gly} \\
 \text{Lys} & \text{Phe} & \text{Pro} & \text{Gly} \\
 (\text{Pro}) & (\text{Pro}) & (\text{Phe}) & (\text{Lys}) \\
 (\text{Gly}) & (\text{Gly}) & (\text{Gly}) & (\text{Pro})
 \end{array}$$

, где первые 4 элемента u_1 соответствуют перекрытиям для 3-х позиций (в скобках указаны альтернативные варианты ,см. столбец m_{3-1} из табл.1), последующие 4 – перекрытиям для первых позиций (см. столбец m_{1-3}). Формальная подстановка базовых кодонов из (5) в (6) приводит кодировки всех 4-х аминокислот к тому, что они становятся неоднозначными по 1-й и 3-ей позициям. Для уточнения приведем вывод всего для двух аминокислот из (5) – Lys и связанной с ней согласно первого перекрытия из (6) – Gly. Согласно сказанному имеем: Lys должна кодироваться набором триплетов X_1aX , а Gly - X_1dN , где $X_1: a, d, c$; $X: a,d$. Тогда имеет место перекрытие Lys с Gly по двум позициям, что невозможно согласно табл.1. Это также не допускает

Таблица 2. Шаги 1-9 процесса расчета генетического кода.

В последней строке указывается число вычисленных кодонов после данного шага. Обозначения X, Y, M, N указаны в подписи к табл. 1.

	1	2	3	4	5	6	7	8	9
Met							abd		
Trp				bdd					
Phe	BBB	BBY							
Tyr						baY			
His							caY		
Asn						aaY			
Asp							daY		
Cys				bdY					
Gln			caX						
Lys	aaa	aaX							
Glu			daX						
Ile							abM		
Val			dbN						
Pro	ccc	ccN							
Thr					acN				
Ala			dcN						
Gly	ddd	ddN							
Ser								bca bcc ddY	bcd bcb
Leu								cbX cbb bbX	cbc
Arg								cdN adX	
Σ	4	12	24	27	31	35	43	58	61

две другие возможности: Lys не может кодироваться набором триплетов X_1aa , если Gly - ddN, а также Lys не может кодироваться набором триплетов aaX , если Gly - X_1dd . Остаются еще две возможности: и Lys и Gly кодируются неоднозначными кодонами по одним и тем же позициям. Случай кодировок Lys: X_1aa , Gly: X_1dd невозможен, так как при этом не может быть удовлетворено условие из табл.1: число m_{23} для Gly равно 5 согласно табл.1. (а для подобной кодировки Gly может быть максимум -4). Поэтому остается единственный из возможных вариантов для рассматриваемых кодировок, когда имеют место неоднозначности в третьей позиции. Аналогично можно установить кодировки для оставшейся пары Phe и Pro. В итоге получим:

$$\text{Lys: } aaX, \text{ Phe: } vvY, \text{ Pro: } ccN, \text{ Gly: } ddN, \quad (7)$$

где Y: b,c.

Из табл.1 следует, что значение m_{12} не превышает числа 4 для аминокислот из (1) с номерами от 1 до 17. Число 4 означает, что первые и вторые позиции могут быть однозначными, чего никак нельзя сказать о m_{12} для Ser, Leu, Arg, для которых подобные значения равны 7,8,7 соответственно. В связи с этим на шагах 3- 7 будут рассматриваться только аминокислоты из (1) с номерами до 17. Отметим, что расчет для всех аминокислот из (1) проводится по методу, который будет опубликован в дальнейшем. Главную роль при этом играют элементы множества U_3 , в каждом из которых кодировки двух аминокислот, а именно A_{ma_1} и A_{ma_2} из рис.1А считаются известными.

Шаг 3. Во множестве U_3 выделим все u_3 , в каждом из которых присутствуют по две аминокислоты из набора (5) и на основании кодировок (7) вычислим соответствующие н.п.:

	Lys	Lys	Phe	Pro	
	Gln	Glu	Val	Ala	
	Pro	Gly	Gly	Gly	(8)
н.п.	ссаа	ддаа	ддвв	ддсс	

откуда имеем однозначные кодировки для 4-х аминокислот: Gln, Glu, Val, Ala, а с учетом m_{3-1} в табл.1 находим:

$$\text{Gln:caX, Glu:daX, Val:dvN, Ala:dcN,} \quad (9)$$

Шаг 4. Во множестве U_3 выделим все u_3 , в каждом из которых присутствуют по две аминокислоты из наборов (5) и (9). Имеем:

$$\begin{array}{ccc} \text{Gly} & \text{Val} & \text{Ala} \\ \text{Trp} & \text{Cys} & \text{Cys} \\ \text{Val} & \text{Val} & \text{Val} \end{array} \quad (10)$$

на основе которых получим

$$\text{Trp: bdd, Cys: vdY} \quad (11)$$

Шаг 5. Во множестве U_3 нет элементов, у которых присутствуют по две аминокислоты из наборов (7), (9), (11). Обратимся ко множеству U_2 и рассмотрим одну из возможностей. На основе двух элементов u_2 :

$$\begin{array}{cc} \text{Thr} & \text{Thr} \\ \text{Gln} & \text{Pro} \end{array} \quad (12)$$

Получим две кодировки для Thr: xca, xcc, где x- пока неизвестно. Для поиска всех кодировок для Thr, обратимся к ее значениям m_{1-3} и m_{3-1} . Из равенства $m_{1-3} = 3$ (Lys, Pro, Gly) следует, что x может принимать не более двух значений: a, d, т.к. в перечисленном наборе отсутствует Phe. Однако значение d невозможно, т.к. dca и dcc являются кодировками Ala. Кроме того, из равенства $m_{3-1}=4$ (это означает, что первые позиции кодировок 4-х аминокислот: Lys, Phe, Pro, Gly перекрываются с кодировками третьей позиции Thr) следует, что третья позиция в кодировках Thr равна N. Таким образом, имеем:

$$\text{Thr: acN.} \quad (13)$$

Шаг 6. Во множестве U_3 выделим все u_3 , содержащих по две аминокислоты, с указанными выше кодировками. Из перекрытий:

$$\begin{array}{cc} \text{Thr} & \text{Thr} \\ \text{Tyr} & \text{Asn} \\ \text{Val} & \text{Gln} \end{array} \quad (14)$$

имеем две кодировки для Tyr: bac, Asn: aac. С учетом данных из столбца m_{3-1} окончательно находим:

$$\text{Tyr: baY, Asn: aaY.} \quad (15)$$

Шаг 7. Обратимся к аналогичным перекрытиям u_3 с учетом двух аминокислот -из (15). Их оказалось 4:

$$\begin{array}{cccc} \text{Trp} & \text{Thr} & \text{Thr} & \text{Phe} \\ \text{Met} & \text{His} & \text{Asp} & \text{Ile} \\ \text{Tyr} & \text{Pro} & \text{Gly} & \text{Tyr} \end{array} \quad (16)$$

с учетом данных из столбца m_{3-1} окончательно находим:

$$\text{Met: abd, His: caY, Asp: daY, Ile: abM,} \quad (17)$$

где M: a, b, c.

Этим шагом завершается поиск кодировок- их оказалось 43- для первых 17-ти аминокислот из табл.1; они даны в столбце K этой таблицы.

Шаг 8. Во множестве U_3 выделим все u_3 , содержащие по одной аминокислоте из набора Ser, Leu, Arg и по две из 17 аминокислот из указанного выше набора. Имеем:

$$\begin{array}{cccc} \text{Gln} & \text{Pro} & \text{Ala} & \text{Val} \\ \text{Ser} & \text{Ser} & \text{Ser} & \text{Ser} \\ \text{Phe} & \text{Ile} & \text{Gln} & \text{Gln} \\ \\ \text{Tyr} & \text{Phe} & \text{Trp} & \text{Tyr} & \text{Trp} \\ \text{Leu} & \text{Leu} & \text{Leu} & \text{Leu} & \text{Leu} \\ \text{Thr} & \text{Ala} & \text{Pro} & \text{Val} & \text{Val} \end{array} \quad (18)$$

$$\begin{array}{cccccc} \text{Glu} & \text{Val} & \text{Ala} & \text{Gly} & \text{Gly} & \text{Gly} \\ \text{Arg} & \text{Arg} & \text{Arg} & \text{Arg} & \text{Arg} & \text{Arg} \\ \text{Lys} & \text{Ala} & \text{Ala} & \text{Pro} & \text{Thr} & \text{Lys} \end{array}$$

Из этих перекрытий находим следующие кодировки:

$$\text{Ser: bca, bcc, adY; Leu: cbX, cbb, bbX; Arg: cdN, adX.} \quad (19)$$

Шаг 9. Во множестве U_3 выделим u_3 , содержащие по две аминокислоты из набора Ser, Leu, Arg и дающие решения, отличные от (19). Имеем:

$$\begin{array}{ccc}
 \text{Arg} & \text{Leu} & \text{Ser} \\
 \text{Ser} & \text{Ser} & \text{Leu} \\
 \text{Val} & \text{Ile} & \text{Pro}
 \end{array} \quad (20)$$

Из первого и второго u_3 находим: Ser: bcd, bcb, а из третьего - Leu: cbc. Окончательные кодировки для Ser, Leu, Arg представлены в столбце К табл.1. А полное число смысловых триплетов для всех 20-и аминокислот из этого столбца равно 61. Три триплета: baX, bda не определяются при использовании множеств U_1, U_2, U_3 ; они дополняют полное число триплетов до 64.

При переходе от набора букв a, b, c, d к каноническим нуклеотидам А, С, Т, G можно получить 24 подобных генетических кода. Всего один из них – стандартный (см. табл.1 из [8]), при a=A, b =Т, c =С, d=G, а триплеты baX, bda становятся ТАА, TAG, TGA; они играют роль терминаторных кодонов – кодонов останавливающих белковый синтез.

Краткое изложение данной работы представлено в [17].

Список литературы

1. Козлов Н.Н. Об особом способе записи генетической информации. ДАН. 1994. Т. 337. № 1. С. 158-161.
2. Козлов Н.Н. Молчащие мутации в области перекрывания генов. ДАН. 1996. Т. 350. № 5. С. 699-703.
3. Козлов Н.Н. Перекрывающиеся гены и генетический код. ДАН. 1997. Т. 355. № 6. С. 830-833.
4. Козлов Н.Н. О востребованности каждого из 64 кодонов в генетических перекрытиях. ДАН. 1999. Т. 367. № 4. С. 544-547.
5. Козлов Н.Н. К вопросу о произвольности «выбора» генетического кода. ДАН. 1999. Т. 369. № 4. С. 553-556.
6. Козлов Н.Н. Анализ полного множества перекрывающихся генов. ДАН. 2000. Т. 373. № 1. С. 108-111.
7. Козлов Н.Н. Перекрывающиеся гены и вариабельность генетического кода. ДАН. 2000. Т. 375. № 6. С. 824-827.
8. Козлов Н.Н. Теорема для генетического кода. ДАН 2002. Т. 382, № 5. С. 593-597.
9. Козлов Н.Н. Применение теоремы для генетического кода. ДАН. 2004. Т. 396. № 6. С. 740-745.
10. Козлов Н.Н. Математический анализ структурных генов. ДАН 2007. Т.412. № 5. С. 610-613.
11. Козлов Н.Н. Математический анализ девиантности генетического кода. ДАН 2007. Т. 415. №. 4, с .441-445.
12. Козлов Н.Н. Интегральные характеристики генетического кода. ДАН 2007. Т. 417. №1. С. 30-33.
13. Козлов Н.Н., Грязнов С.С. Некоторые расчетные характеристики больших геномов. ДАН 2007. Т. 417. № 6. С.732-737.
14. Козлов Н.Н. Множества, порождаемые генетическим кодом. ДАН 2008. Т.423. №. 3. С.295- 299.
15. Козлов Н.Н. Неоднозначности во множествах, порождаемых генетическим кодом. ДАН 2010. Т.432. № 2. С.157- 161.
16. Козлов Н.Н. Математический анализ генетического кода», М., БИНОМ, 2010, 223 с.
17. Козлов Н.Н. Расчет генетического кода. ДАН 2010. Т.433. №.2.

ПРИЛОЖЕНИЕ

Ниже приводится полный перечень элементов множеств U_2 и U_3 .
На стр.16-18 приводятся все 160 элементов множества U_2 .

На стр.19-23 приводятся все 307 элементов множества U_3 .
Элементы приводятся в том представлении, которое используется в данной задаче. Каждый из элементов состоит из трех строк: верхней, средней и нижней. Элементы множества упорядочены по средней строке в порядке нумерации аминокислот, представленной в табл.1.

1	Met	Met	Met	Met
Met	Tyr	His	Asn	Asp
(4+2)	1	2	3	4
	Met	Met		
	Trp	Cys		
	5	6		
2	Trp	Trp	Trp	
Trp	Met	Val	Leu	
(3+1)	7	8	9	
	Trp			
	Gly			
	10			
3	Phe	Phe	Phe	Phe
Phe	Phe	Ile	Val	Leu
(4+3)	11	12	13	14
	Phe	Phe	Phe	
	Phe	Ser	Leu	
	15	16	17	
4	Tyr	Tyr	Tyr	
Tyr	Ile	Val	Leu	
(3+3)	18	19	20	
	Tyr	Tyr	Tyr	
	Met	Ile	Thr	
	21	22	23	
5	His	His	His	His
His	Pro	Thr	Ala	Ser
(4+3)	24	25	26	27
	His	His	His	
	Met	Ile	Thr	
	28	29	30	
6	Asn	Asn	Asn	
Asn	Gln	Lys	Glu	
(3+3)	31	32	33	
	Asn	Asn	Asn	
	Met	Ile	Thr	
	34	35	36	
7	Asp	Asp		
Asp	Gly	Arg		
(2+3)	37	38		
	Asp	Asp	Asp	
	Met	Ile	Thr	
	39	40	41	

8	Cys	Cys	Cys			
Cys	Met	Val	Leu			
(3+2)	42	43	44			
	Cys	Cys				
	Val	Ala				
	45	46				
9	Gln	Gln	Gln	Gln		
Gln	Pro	Thr	Ala	Ser		
(4+4)	47	48	49	50		
	Gln	Gln	Gln	Gln		
	Asn	Lys	Ser	Arg		
	51	52	53	54		
10	Lys	Lys	Lys			
Lys	Gln	Lys	Glu			
(3+4)	55	56	57			
	Lys	Lys	Lys	Lys		
	Asn	Lys	Ser	Arg		
	58	59	60	61		
11	Glu	Glu				
Glu	Gly	Arg				
(2+4)	62	63				
	Glu	Glu	Glu	Glu		
	Asn	Lys	Ser	Arg		
	64	65	66	67		
12	Ile	Ile	Ile	Ile		
Ile	Tyr	His	Asn	Asp		
(4+4)	68	69	70	71		
	Ile	Ile	Ile	Ile		
	Phe	Tyr	Ser	Leu		
	72	73	74	75		
13	Val	Val	Val	Val		
Val	Cys	Gly	Ser	Arg		
(4+6)	76	77	78	79		
	Val	Val	Val	Val	Val	Val
	Trp	Phe	Tyr	Cys	Ser	Leu
	80	81	82	83	84	85
14	Pro	Pro	Pro	Pro		
Pro	Pro	Thr	Ala	Ser		
(4+5)	86	87	88	89		
	Pro	Pro	Pro	Pro	Pro	
	His	Gln	Pro	Leu	Arg	
	90	91	92	93	94	

15	Thr	Thr	Thr	Thr				
Thr	Tyr	His	Asn	Asp				
(4+5)	95	96	97	98				
	Thr	Thr	Thr	Thr	Thr			
	His	Gln	Pro	Leu	Arg			
	99	100	101	102	103			
16	Ala	Ala	Ala	Ala				
Ala	Cys	Gly	Ser	Arg				
(4+5)	104	105	106	107				
	Ala	Ala	Ala	Ala	Ala			
	His	Gln	Pro	Leu	Arg			
	108	109	110	111	112			
17	Gly	Gly	Gly					
Gly	Trp	Gly	Arg					
(3+5)	113	114	115					
	Gly	Gly	Gly	Gly	Gly			
	Asp	Glu	Val	Ala	Gly			
	116	117	118	119	120			
18	Ser							
Ser	Phe	Gln	Lys	Glu	Ile	Val	Leu	
(7+7)	121	122	123	124	125	126	127	
	Ser							
	His	Gln	Val	Pro	Ala	Leu	Arg	
	128	129	130	131	132	133	134	
19	Leu							
Leu	Phe	Ile	Val	Pro	Thr	Ala	Ser	Leu
(8+6)	135	136	137	138	139	140	141	142
	Leu	Leu	Leu	Leu	Leu	Leu		
	Trp	Phe	Tyr	Cys	Ser	Leu		
	143	144	145	146	147	148		
20	Arg							
Arg	Gln	Lys	Glu	Pro	Thr	Ala	Ser	
(7+5)	149	150	151	152	153	154	155	
	Arg	Arg	Arg	Arg	Arg			
	Asp	Glu	Val	Ala	Gly			
	156	157	158	159	160			

Trp	Trp	Trp	Trp	Cys	Cys	Cys	Cys
Met							
Tyr	His	Asn	Asp	Tyr	His	Asn	Asp
1	2	3	4	5	6	7	8

Gly	Gly	Gly
Trp	Trp	Trp
Met	Val	Leu
9	10	11

Phe	Phe	Phe	Phe	Ser	Ser
Phe	Phe	Phe	Phe	Phe	Phe
Phe	Ile	Val	Leu	Phe	Ile
12	13	14	15	16	17

Ser	Ser	Leu	Leu	Leu	Leu
Phe	Phe	Phe	Phe	Phe	Phe
Val	Leu	Phe	Ile	Val	Leu
18	19	20	21	22	23

Met	Met	Met	Ile	Ile	Ile	Thr	Thr	Thr
Tyr								
Ile	Val	Leu	Ile	Val	Leu	Ile	Val	Leu
24	25	26	27	28	29	30	31	32

Met	Met	Met	Met	Ile	Ile
His	His	His	His	His	His
Pro	Thr	Ala	Ser	Pro	Thr
33	34	35	36	37	38

Ile	Ile	Thr	Thr	Thr	Thr
His	His	His	His	His	His
Ala	Ser	Pro	Thr	Ala	Ser
39	40	41	42	43	44

Met	Met	Met	Ile	Ile	Ile
Asn	Asn	Asn	Asn	Asn	Asn
Gln	Lys	Glu	Gln	Lys	Glu
45	46	47	48	49	50

Thr	Thr	Thr
Asn	Asn	Asn
Gln	Lys	Glu
51	52	53

Met	Met	Ile	Ile	Thr	Thr
Asp	Asp	Asp	Asp	Asp	Asp
Gly	Arg	Gly	Arg	Gly	Arg
54	55	56	57	58	59

Val	Val	Val	Ala	Ala	Ala
Cys	Cys	Cys	Cys	Cys	Cys
Met	Val	Leu	Met	Val	Leu
60	61	62	63	64	65

Asn	Asn	Asn	Asn	Lys	Lys	Lys	Lys
Gln							
Pro	Thr	Ala	Ser	Pro	Thr	Ala	Ser
66	67	68	69	70	71	72	73

Ser	Ser	Ser	Ser	Arg	Arg	Arg	Arg
Gln							
Pro	Thr	Ala	Ser	Pro	Thr	Ala	Ser
74	75	76	77	78	79	80	81

Asn	Asn	Asn	Lys	Lys	Lys
Lys	Lys	Lys	Lys	Lys	Lys
Gln	Lys	Glu	Gln	Lys	Glu
82	83	84	85	86	87

Ser	Ser	Ser	Arg	Arg	Arg
Lys	Lys	Lys	Lys	Lys	Lys
Gln	Lys	Glu	Gln	Lys	Glu
88	89	90	91	92	93

Asn	Asn	Lys	Lys	Ser	Ser	Arg	Arg
Glu							
Gly	Arg	Gly	Arg	Gly	Arg	Gly	Arg
94	95	96	97	98	99	100	101

Phe	Phe	Phe	Phe	Tyr	Tyr	Tyr	Tyr
Ile							
Tyr	His	Asn	Asp	Tyr	His	Asn	Asp
102	103	104	105	106	107	108	109

Ser	Ser	Ser	Ser	Leu	Leu	Leu	Leu
Ile							
Tyr	His	Asn	Asp	Tyr	His	Asn	Asp
110	111	112	113	114	115	116	117

Trp	Trp	Trp	Trp	Phe	Phe	Phe	Phe
Val							
Cys	Gly	Ser	Arg	Cys	Gly	Ser	Arg
118	119	120	121	122	123	124	125

Tyr	Tyr	Tyr	Tyr	Cys	Cys	Cys	Cys
Val							
Cys	Gly	Ser	Arg	Cys	Gly	Ser	Arg
126	127	128	129	130	131	132	133

Ser	Ser	Ser	Ser	Leu	Leu	Leu	Leu
Val							
Cys	Gly	Ser	Arg	Cys	Gly	Ser	Arg
134	135	136	137	138	139	140	141

His	His	His	His	Gln	Gln	Gln	Gln
Pro							
Pro	Thr	Ala	Ser	Pro	Thr	Ala	Ser
142	143	144	145	146	147	148	149

Pro	Pro	Pro	Pro	Leu	Leu	Leu	Leu
Pro							
Pro	Thr	Ala	Ser	Pro	Thr	Ala	Ser
150	151	152	153	154	155	156	157

Arg	Arg	Arg	Arg
Pro	Pro	Pro	Pro
Pro	Thr	Ala	Ser
158	159	160	161

His	His	His	His	Gln	Gln	Gln	Gln
Thr							
Tyr	His	Asn	Asp	Tyr	His	Asn	Asp
162	163	164	165	166	167	168	169

Pro	Pro	Pro	Pro	Leu	Leu	Leu	Leu
Thr							
Tyr	His	Asn	Asp	Tyr	His	Asn	Asp
170	171	172	173	174	175	176	177

Arg	Arg	Arg	Arg
Thr	Thr	Thr	Thr
Tyr	His	Asn	Asp
178	179	180	181

His	His	His	His	Gln	Gln	Gln	Gln
Ala							
Cys	Gly	Ser	Arg	Cys	Gly	Ser	Arg
182	183	184	185	186	187	188	189

Pro	Pro	Pro	Pro	Leu	Leu	Leu	Leu
Ala							
Cys	Gly	Ser	Arg	Cys	Gly	Ser	Arg
190	191	192	193	194	195	196	197

Arg	Arg	Arg	Arg
Ala	Ala	Ala	Ala
Cys	Gly	Ser	Arg
198	199	200	201

Asp	Asp	Asp	Glu	Glu	Glu	Val	Val	Val
Gly								
Trp	Gly	Arg	Trp	Gly	Arg	Trp	Gly	Arg
202	203	204	205	206	207	208	209	210

Ala	Ala	Ala	Gly	Gly	Gly
Gly	Gly	Gly	Gly	Gly	Gly
Trp	Gly	Arg	Trp	Gly	Arg
211	212	213	214	215	216

His	His	His	His	Gln	Gln	Gln	Gln
Ser							
Phe	Ile	Val	Leu	Phe	Ile	Val	Leu
217	218	219	220	221	222	223	224

Pro	Pro	Pro	Pro	Leu	Leu	Leu	Leu
Ser							
Phe	Ile	Val	Leu	Phe	Ile	Val	Leu
225	226	227	228	229	230	231	232

Arg	Arg	Arg	Arg	Val	Val	Val	Ala
Ser							
Phe	Ile	Val	Leu	Gln	Lys	Glu	Gln
233	234	235	236	237	238	239	240

Ala	Ala
Ser	Ser
Lys	Glu
241	242

Trp	Trp	Trp	Trp	Phe	Phe	Phe	Phe
Leu							
Pro	Thr	Ala	Ser	Pro	Thr	Ala	Ser
243	244	245	246	247	248	249	250

Tyr	Tyr	Tyr	Tyr	Cys	Cys	Cys	Cys
Leu							
Pro	Thr	Ala	Ser	Pro	Thr	Ala	Ser
251	252	253	254	255	256	257	258

Ser	Ser	Ser	Ser	Leu	Leu	Leu	Leu
Leu							
Pro	Thr	Ala	Ser	Pro	Thr	Ala	Ser
259	260	261	262	263	264	265	266

Trp	Trp	Trp	Trp	Tyr	Tyr	Tyr	Tyr
Leu							
Phe	Ile	Val	Leu	Phe	Ile	Val	Leu
267	268	269	270	271	272	273	274

Cys	Cys	Cys	Cys
Leu	Leu	Leu	Leu
Phe	Ile	Val	Leu
275	276	277	278

Asp	Asp	Asp	Asp	Glu	Glu	Glu	Glu
Arg							
Pro	Thr	Ala	Ser	Pro	Thr	Ala	Ser
279	280	281	282	283	284	285	286

Val	Val	Val	Val	Ala	Ala	Ala	Ala
Arg							
Pro	Thr	Ala	Ser	Pro	Thr	Ala	Ser
287	288	289	290	291	292	293	294

Gly	Gly	Gly	Gly	Asp	Asp	Asp
Arg						
Pro	Thr	Ala	Ser	Gln	Lys	Glu
295	296	297	298	299	300	301

Glu	Glu	Glu	Gly	Gly	Gly
Arg	Arg	Arg	Arg	Arg	Arg
Gln	Lys	Glu	Gln	Lys	Glu
302	303	304	305	306	307