Российская Академия Наук
Ордена Ленина
Институт прикладной математики
им. М.В.Келдыша

Н.Н.Козлов

ЭЛЕМЕНТАРНЫЕ ГЕНЕТИЧЕСКИЕ ПЕРЕКРЫТИЯ

Москва, 2004

Н.Н.Козлов ЭЛЕМЕНТАРНЫЕ ГЕНЕТИЧЕСКИЕ ПЕРЕКРЫТИЯ. Препринт ИПМ им. М.В.Келдыша, РАН, М., 2004.

Рассматривается множество элементов, порождаемых генетическим кодом. Множество вычисляется для необычных способов записи генетической информации, называемой перекрывающимися генами, когда один и тот же участок ДНК может кодировать две белковые последовательности. Вводится понятие элементарного перекрытия, которое соответствует перекрытию для одиночных аминокислот. Построено пять множеств элементарных перекрытий, каждое из которых соответствует одному из пяти способов генетических перекрытий, обнаруженных экспериментально. Анализируется структура множеств элементарных перекрытий и их свойства. Дается подробное и сжатое представление множеств элементарных перекрытий. Рассматриваются вопросы применения множеств.

Ключевые слова: перекрывающиеся гены, генетический код, вырожденность кода, девиантные коды, происхождение кода, эволюция кода.

N.N.Kozlov. Elementary genetic overlappings.. Preprint KIAM RAN, Moscow, 2004.

The set of elements produced by a genetic code is examined. The set is calculated for unusual ways of record of the genetic information named as overlapped genes, when the same site of DNA can code two protein sequences. The concept of elementary overlapping is entered which corresponds to overlapping for single amino acids. Five sets of elementary overlappings are constructed, each of which corresponds to one of five ways of the genetic overlappings which have been found out experimentally. The structure of sets of elementary overlappings and their properties is analyzed. The detailed and compressed representation of sets of elementary overlappings is given. The questions of application of sets are examined.

Key words: overlapping genes, genetic code, degenerated code, deviante code, origin code, evolution code.

Работа выполнена при финансовой поддержке Программы фундаментальных исследований Президиума РАН «Параллельные вычисления на многопроцессорных вычислительных системах», Российского фонда фундаментальных исследований (коды проектов 04-01-00320 и 02-07-90027), Минпромнауки (Госконтракт № 37.011.11.0012), а также гранта ведущих научных школ (НШ-2003.2003.1)

Введение

В последние годы был проведен математический анализ структуры генетического кода [1]. Этот анализ основывался на исследовании некоторых необычных способов записи структурных генов. Структурные гены - гены кодирующие белки, представлены на одной из цепей двухцепочечной ДНК и записываются 4-х буквенным алфавитом: A,T,C,G - это нуклеотиды аденин, тимин, цитозин и гуанин соответственно. Устойчивость ДНК поддерживается комплементарными связями, когда A в одной цепи ДНК связывается только с T из другой цепи, а C - только с G. Эксперименты показали, что кодировка белков осуществляется через тройки нуклеотидов: AAA,AAT,AAC,...GGG, которые называются триплетами или кодонами и являются единицами генетического кода. Всего имеем 64 триплета. Оказалось, что лишь 61 триплет из них кодирует одну из 20-и белковых единиц - аминокислот, а триплеты TAA,TAG,TGA не кодируют никакой аминокислоты. Причем оказалось, что некоторые аминокислоты могут кодироваться одним триплетом, а другие большим числом триплетов, вплоть до 6-и. Такая неоднозначность была названа вырожденностью кода. Сам генетический код (см. табл. 1 из [1]) может рассматриваться как оператор преобразования между генами и белками или точнее между кодонными семействами и 20-ю аминокислотами. Перечень аминокислот в стандартом виде (трехбуквенные сокращения) приводится ниже в табл. 1. Указанный код называют стандартным - K^0 (первоначальное название - универсальный), которым записано подавляющее большинство структурных генов, известных в настоящее время. Через 13 лет (в 1979г.) после того, как этот код был окончательно установлен, было показано, что код в природе не один. На сегодня известно уже более 10 нестандартных кодов для некоторых организмов. Нами были проанализированы ряд таких кодов [2].

Необычный способ записи генов, о котором говорилось выше, состоит в том, что один и тот же участок цепи ДНК, кодирующий белок, может читаться со сдвигом фазы на +1 либо -1 нуклеотид. Иными словами один и тот же указанный участок может кодировать два и более белков. Такие гены были названы перекрывающимися. Отметим, что как показывают эксперименты, такое чтение оказывается разрешенным лишь в некоторых случаях, а в подавляющем большинстве случаев существует запрет на указанные альтернативные чтения. Этот запрет состоит в том, что указанные сдвиги приводят к совершенно иным последовательностям кодонов отличным от исходной последовательности (когда сдвигов нет). Но было установлено, что в подобных альтернативных последовательностях непременно возникают какие-либо кодоны из трех: TAA, TAG, TGA указанных выше (так устроен ген кодирующий белок, или так выбраны соответствующие кодировки аминокислот вследствие вырожденности кода). Роль названных трех кодонов одинакова - они останавливают (блокируют) белковый синтез, ко-

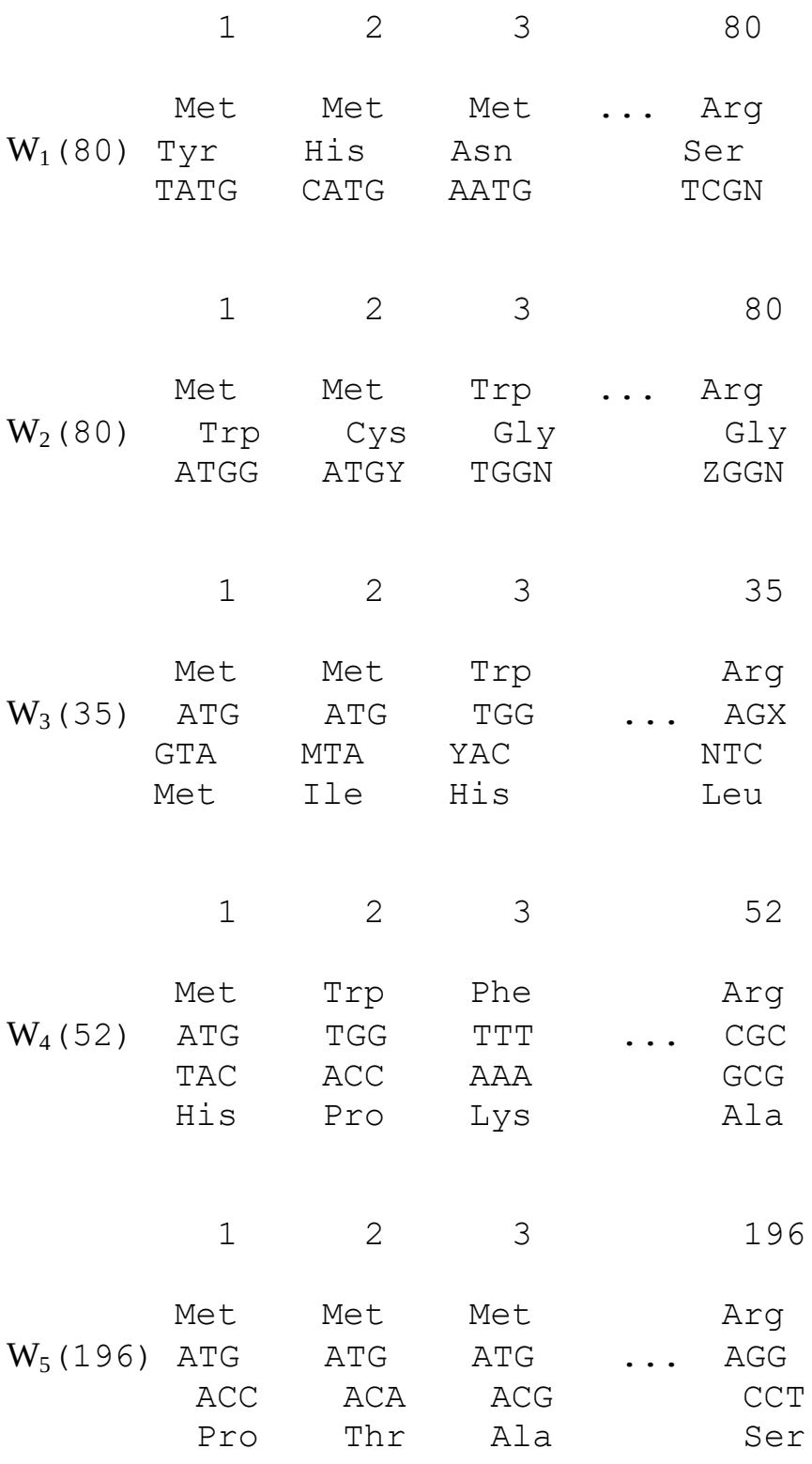

Рис. 1. Описание структуры множеств W_1-W_5.
Приведены по 4-е э.п. в каждом из множеств.

торый происходит (по тексту гена). Иными словами белок при альтернативном чтении не синтезируется. Оказалось, что лишь для перекрывающихся генов такого запрета не существует. Впервые этот эффект был установлен в 1976 году при чтении первого целого генома - вируса бактерии ФХ 174 [3]. ДНК такого вируса оказалась кольцевой и одноцепочечной и содержащей 5386 нуклеотидов. Одно из самых длинных перекрытий, обнаруженных к настоящему времени, относится к ДНК вируса GSHV [4] и содержит около 1300 нуклеотидов, а полный размер всех перекрытий равен 1704 нуклеотида или более половины от размера всей ДНК (одна цепь содержит - 3311 нуклеотида).

Названные два генома содержат случаи перекрытий генов, принадлежащих одной цепи ДНК. Таких случаев 2: сдвиг на +1 либо -1 нуклеотид относительно исходного гена. Позднее были установлены перекрытия двух генов, принадлежащих различным цепям ДНК, которые называются + цепью и - цепью ДНК. При этом следует отметить, что чтение гена (и последовательностей триплетов) происходит слева направо для + цепи ДНК и справа налево для - цепи ДНК. Случаев перекрывания при этом будет 3: сдвиг на +1, 0 либо -1 нуклеотид гена из - цепи относительно гена из + цепи. Таким образом, полное число случаев перекрытий пар генов равно 5. Экспериментальные данные по всем таким случаям представлены на рис. 1 из [1].

2. Элементарные перекрытия.

Основные результаты из [1, 2] были получены на основе математического анализа множеств элементарных перекрытий - э.п. Таких - множеств 5: W_1-W_5, каждое из которых соответствует одному из 5-и указанных случаев перекрывания. Э.п. - это перекрытие соответствующее одиночным аминокислотам. Поскольку здесь есть неопределенность: сколько нуклеотидов может перекрываться 1, 2 либо 3 (если возможно) нуклеотида, то под э.п. понимаем перекрывание максимально возможного числа нуклеотидов: это 3 нуклеотида для случая 4 и 2 для всех остальных случаев перекрытий. На рис.1 дано краткое представление э.п.: представлены лишь по 4 э.п. для каждого из множеств W_1-W_5. Опишем первые э.п. в этих множествах. Для W_1 первое э.п. соответствует перекрытию кодона `Met` и кодона `Tyr`, общая пара нуклеотидов `AT`, сдвиг между кодонами равен -1 нуклеотид. Для W_2 первое э.п. соответствует перекрытию `Met` и кодона `Trp`, общая пара нуклеотидов `TG`, сдвиг между кодонами равен +1 нуклеотид. В отличие от W_1 и W_2 э.п. из W_3-W_5 соответствуют разным цепям ДНК.

Верхний кодон в этих э.п. соответствует + цепи ДНК и чтение кодона идет слева направо, а нижний кодон соответствует - цепи ДНК и чтение кодона идет справа налево. Для W_3 первое э.п. соответствует перекрытию кодона `ATG(Met)` и кодона `Met`, который справа налево читается как

Таблица 1
Число элементарных перекрытий (э.п.)
в 20-и подмножествах каждого из множеств W_1-W_5

		W_1	W_2	W_3	W_4	W_5	$\sum$
1	Met	4	2	2	1	4	13
2	Trp	3	1	2	1	4	11
3	Phe	4	3	2	2	5	16
4	Tyr	3	3	1	2	8	17
5	His	4	3	2	2	8	19
6	Asn	3	3	2	2	8	18
7	Asp	2	3	1	2	8	16
8	Cys	3	2	2	2	8	17
9	Gln	4	4	2	1	8	19
10	Lys	3	4	2	2	8	19
11	Glu	2	4	1	2	8	17
12	Ile	4	4	2	3	8	21
13	Val	4	6	1	4	12	27
14	Pro	4	5	1	3	13	26
15	Thr	4	5	1	4	13	27
16	Ala	4	5	1	4	13	27
17	Gly	3	5	1	4	15	28
18	Ser	7	7	3	4	18	39
19	Leu	8	6	4	3	12	33
20	Arg	7	5	2	4	15	33
$\sum$		80	80	35	52	196	443
$\overline{\mu}$		4	4	1.75	2.6	9.8	~22
$\mu_{\min}$		2	1	1	1	4	11
$\mu_{\max}$		8	7	4	4	18	39

ATG. Сдвиг между кодонами, принадлежащих разным цепям ДНК, составляет -1 нуклеотид. Нуклеотиды из пары AT в + цепи комплементарно связаны с нуклеотидами TA из - цепи ДНК: это связи AT и TA. Для W_4 первое э.п. соответствует перекрытию кодона Met и кодона His(CAT). Кодоны берутся из разных цепей ДНК и сдвиг между кодонами отсутствует. Перекрытию соответствуют 3 комплементарные связи: AT,TA,GC. Для W_5 первое э.п. соответствует перекрытию кодона Met и кодона Pro(CCA). Сдвиг между кодонами из разных цепей ДНК составляет +1 нуклеотид, в перекрытии 2 комплементарные связи TA,GC.

Полный набор э.п. в каждом из множеств указывается в скобках: для W_1 и W_2 это 80, для W_3 - 35, для W_4 - 52, для W_5 - 196. На стр. 13-27 приводится полный перечень этих множеств, в которых приведены ряд сокращений. Это N:A,C,T,G; Y:T,C; X:A,G; Z:A,C; I:T,G. Порядок следования кодонов соответствует левому столбцу таблицы 1: Met, Trp, Phe, … Arg. Отметим, что в полном представлении множества W_5 фактическое число э.п. на 5 превышает - 196. Речь идет о некоторых перекрытиях для трех аминокислот: Gly, Ser, Arg. Для Gly имеем два э.п. с номерами 144.1 и 144.2 у которых как в верхней, так и в нижней позициях одни и те же аминокислоты: Gly и Ser соответственно. Такое оказалось возможным благодаря нерегулярности кодонного семейства Ser: в одном из названных э.п. использован кодон Ser из набора TCN, а в другом - из набора AGY. Аналогичная картина наблюдается в трех парах э.п. : 156.1 и 156.2, 157.1 и 157.2, 164.1 и 164.2, у каждой из которых участвуют кодоны семейства Ser, а также кодон из наборов Arg: CGN или AGX. Кодоны из наборов Ser и Arg присутствуют также в паре э.п. 188.1 и 188.2. В связи с такой особенностью первые 3 цифры э.п. в 5-и указанных парах одинаковы и будем полагать (в дальнейшем анализе) что подобные пары соответствуют одному э.п. Это объясняется тем, что главный анализ относится к анализу аминокислот в э.п., а в каждой из названных пар аминокислоты в перекрытиях одинаковы.

3. Анализ множеств элементарных перекрытий.

В полном представлении множеств W_1-W_5 э.п. приведены последовательно в подмножествах, каждое из которых отличается от другого различием аминокислот в верхних позициях. Поэтому число подмножеств в каждом из множеств W_1-W_5 равно 20, а число э.п. в каждом из подмножеств указывается в скобках перед первым э.п. соответствующего подмножества. Эти 100 чисел - μ_{ij} - выделены также в таблице 1. Отметим некоторые особенности множеств W_1-W_5. Все эти множества, кроме W_4, содержат самоперекрытия, это когда одна и та же аминокислота присутствует как в верхней, так и в нижней строке. При этом оказалось, что в самоперекрытиях

участвуют все аминокислоты кроме трех: Trp, Gln, Glu, т.е. 17 аминокислот. Причем это участие может быть в трех множествах: это Pro, Gly, Leu - в W_1, W_2, W_5 либо в двух множествах: это Phe, Lys в W_1, W_2, а также Tyr, Ile, Ala, Arg в W_3, W_5. Кроме того имеют место самоперекрытия для отдельно взятых множеств: это His, Asn, Asp, Cys, Val, Thr, Ser в W_5, а Met в W_3.

Сравним численность всех э.п. для одной и двух цепей ДНК. Формальное сравнение это 160 и 283, а фактическое - это 80 и 283 т.е. число э.п. для двух цепей ДНК более чем в 3,5 раза больше. Анализ этого феномена обсуждается в дальнейших публикациях. Покажем, что множества W_1 и W_2 фактически одинаковы; их отличает только перестановка аминокислот в строках. Например, э.п. с номером 1 из W_1 адекватно э.п. с номером 7 из W_2. Для сравнения укажем, что в множествах W_3-W_5 такого не наблюдается: ни один э.п. в этих множествах не может быть получен с помощью простой перестановки аминокислот в каком-либо э.п., принадлежащем другому множеству. Иными словами W_3-W_5 содержат э.п. по существу различающиеся, в отличие от э.п. из W_1, W_2. Возникает вопрос, зачем тогда нужны 2 множества W_1 и W_2? Дело в том, что происхождение этих множеств связана с решением конкретной задачи (такая задача ставилась для каждого множества W_1-W_5): сформировать на основе множества э.п. все возможные наборы перекрытий или из множеств э.п. построить множества для перекрывания двух, трех и более аминокислот. Оказалось, что при такой сборке решения получились разными. Суть этого различия укажем кратко. При использовании э.п. из W_1 в верхней строке могут быть получены последовательности аминокислот с любыми их сочетаниями [6]. При использовании э.п. из W_2 имеет место ограничение [7]: невозможно получить аминокислотные последовательности, содержащие хотя бы одну из 5-и пар аминокислот: MetMet, MenAsn, MetLys, MetIle, MetThr. Решение указанной задачи сборки э.п. для всех множеств W_1-W_5 составляет основу теоремы для генетического кода [1]. Отметим, что парные генетические перекрытия могут быть исследованы при решении задачи сборки э.п., принадлежащих одному множеству. При изучении перекрытий трех, четырех, пяти и шести генов могут быть использованы суперпозиции э.п. из двух, трех, четырех и пяти различных множеств э.п. соответственно. В этом смысле задача сборки имеет более широкий смысл, чем сборка э.п. из одного множества.

Обратимся к таблице 1, где представлены численные значения э.п. μ_{ij}, где i - номер подмножества, j - номер случая перекрытия (или номера множеств W_1-W_5). Первое, что следует сказать состоит в том, что для всех i, j имеем $\mu_{ij} \geq 1$, т.е. для каждой из 20-и аминокислот существует хотя бы одно э.п. по какому-либо из 5-и способов перекрытия. Таким образом, ни для какого-либо из 100 значений μ_{ij} нет нуля; это важно для кода K^0,

ориентированного на множественные перекрытия [1]. Минимальное значение $\mu = 1$ наблюдается только в множествах W_2, W_3, W_4, для W_1 имеем $\mu_{\min} = 2$, для W_5 - $\mu_{\min} = 4$. Число подмножеств для которых $\mu = 1$ равно 12: одно для W_2, 3 для W_4 и 8 для W_3. Максимальное значение μ соответствует подмножествам *Ser* (W_2, W_5), *Leu* (W_1, W_3). Значение $\mu_{\max} = 4$ для W_4 соответствует 6-и подмножествам: Arg, Ser, Gly, Ala, Thr, Val. Максимальное значение из 100 чисел μ_{ij} соответствует Ser (W_5) и равно 18. Таким образом оказалось, что для K^0 не существует ни одной кодировки, которая содержала бы э.п. со всеми 20-ю аминокислотами. Указанное значение $\mu = 18$ соответствуют э.п. для Ser со всеми аминокислотами, кроме Phe и Ile.

Полное число э.п. для множеств W_1-W_5 равно 443. Среднее значение - 22, минимальное соответствует Trp, и равно 11, а максимальное - Ser, и равно - 39. Представим сжатое описание всех э.п. На рис. 2 по оси абсцисс указывается номер аминокислоты из верхней строки в э.п., по оси ординат - номер аминокислоты из нижней строки в э.п. Число возможных позиций (клеток) равно 400. Ясно, что для 443 э.п. некоторые из позиций должны быть заняты э.п., принадлежащим различным множествам. Такие позиции указаны 16-ю двухзначными числами 12, 14, … 75, позиций подобного рода 113. Кроме того 182 позиции заняты числами 1-5, которые соответствуют э.п. из множеств W_1-W_5. Таким образом, из 400 возможных позиций только 295 (113+182) заняты какими-либо э.п., а оставшиеся 105 позиций являются свободными. Иными словами э.п. из всех множеств W_1-W_5 не содержат все возможные перекрывания для любых пар аминокислот. Максимальное число свободных позиций равно 10 и соответствует Trp, а минимальное число - 0 и соответствует Ser, среднее число подобных позиций чуть более 5-и (105/20). Иными словами каждая аминокислота имеет в среднем 15 э.п. для всех множеств W_1-W_5. Максимально это число только для Ser - 20 или только Ser содержит э.п. с каждой из 20-и аминокислот, если рассматривать все 5 способов перекрываний.

Рассмотрим позиции, которые соответствуют 2-м и более э.п. Имеем 4 позиции, каждая из которых содержит по 4 э.п.: две позиции 65 - это э.п. из множеств с номерами 1, 2, 3, 5 и две позиции 75 - это э.п. из множеств с номерами 1, 2, 4, 5. Кроме того, имееем 82 позиции занятых э.п. из двух множеств; таких пар э.п. всего 7 групп из которых: 10 позиций 12(12 - это э.п. из множеств W_1, W_2), 5 позиций 14, 19 позиций 15, 5 позиций 24, 18 позиций 25, 12 позиций 35, 13 позиций 45. Позиции, занятые тремя э.п. обозначены также двухзначным числом. Номера трех множеств W, которым принадлежат такие э.п. указаны в скобках. Всего подобных позиций 27, они составят тройки э.п. из 7 групп: это 2 позиции 16 (1, 2, 4), 7 позиций 17 (1, 2,5), 1 позиция 18 (1, 3,5), 5 позиций э.п. 19 (1, 4,5), 1 позиция 28 (2, 3, 5), 5 позиций 29 (2, 4, 5) и 6 позиций 39 (3, 4, 5). На рис. 2 можно выде-

лить две квадратные области, все позиции которых содержит э.п. (нет пустых позиций). Это область из 25 позиций для э.п. с номерами 4-8 (Tyr–Cys) и область для э.п. с номерами 13-20 (Val–Arg). Последняя область содержит 64 позиции и соответствует всем э.п. только для аминокислот с кодировками от 4-х до 6-и триплетов; использованы все аминокислоты с подобными кодировками. Первая область принадлежит лишь э.п. для пяти аминокислот (Tyr, His, … Cys) из 9-и возможных, каждая из которых кодируется двумя триплетами.

4. Применение множеств.

1. Множества W_1-W_5 впервые были использованы при доказательстве теоремы для генетического кода [1]. Оказалось, что при использовании K^0 почти для любых последовательностей аминокислот могут существовать перекрытия для каждого из 5-и случаев. Исключение составят лишь последовательности, содержащие хотя бы одну из 16 пар аминокислот, которые были установлены.

2. Для анализа вопроса о произвольности «выбора» генетического кода [1] были рассмотрены ряд гипотетических кодов отклоненных от K^0 вследствие перестановок некоторых кодонов. Выбор этих перестановок не был случайным, а опирался на анализ всех 100 подмножеств для множеств W_1-W_5. Было установлено, что одиночная перестановка может привести к увеличению числа указанных выше пар аминокислот не более чем на 12 (т.е. к увеличению до 28 пар). Анализировались также гипотетические коды с множеством перестановок, для которых указанное число пар может быть увеличено почти на порядок.

3. На основе коррекции множеств W_1-W_5 изучались вопросы применения указанной теоремы для природных нестандартных кодов [2]. Анализ всех таких кодов показал, что они отличаются от K^0 небольшим числом перестановок кодонов. Поэтому для анализа каждого из таких кодов требуется лишь относительно небольшая коррекция множеств W_1-W_5. Такой подход позволил получить решения для ряда нестандартных кодов, в том числе для кода, которым записываются белки в митрохондрии человека [2].

4. Ограниченный набор э.п. был изучен нами ранее при анализе нерегулярностей кодонных семейств Leu, Ser, Arg [7]. При этом рассматривались только перекрытия генов из одной цепи ДНК. Таких э.п. оказалось всего 12 (см. табл. 1 из [7]). Анализ этих э.п. показал их возможную роль в расширении протяженности перекрытий как двух, так и трех генов в одной цепи ДНК (см. рис. 2 из [7]).

Автор благодарен О.Н.Козловой и Т.И.Кругловой за помощь в подготовке препринта.

СПИСОК ЛИТЕРАТУРЫ

1. Н.Н.Козлов. Теорема для генетического кода. ДАН. 2002. Т. 382. № 5. С. 593-597.
2. Н.Н.Козлов. Применение теоремы для генетического кода. ДАН. 2004. Т. 396. № 6. С. 740-745.
3. Sanger F., Coulson A.R., Friedmann et al. // J. Mol. Biol. 1978. V. 125. P. 225-246.
4. Seeger C., Ganem D., Varmus H.E. // J.Virol. 1984. V. 51. P. 367-375.
5. Н.Н.Козлов. К вопросу о произвольности «выбора» генетического кода. ДАН. 1999. Т. 369. № 4. С. 553-556.
6. Н.Н.Козлов. Анализ полного множества перекрывающихся генов. ДАН. 2000. Т. 373. № 1. С. 108-111.
7. Н.Н.Козлов. Перекрывающиеся гены и генетический код. ДАН. 1997. Т. 355. № 6. С. 830-833.

W_1

1	Met	Met	Met	Met
Met	Tyr	His	Asn	Asp
(4)	TATG	CATG	AATG	GATG
	1	2	3	4
2	Trp	Trp	Trp	
Trp	Met	Val	Leu	
(3)	ATGG	GTGG	YTGG	
	5	6	7	
3	Phe	Phe	Phe	Phe
Phe	Phe	Ile	Val	Leu
(4)	TTTY	ATTY	GTTY	CTTY
	8	9	10	11
4	Tyr	Tyr	Tyr	
Tyr	Ile	Val	Leu	
(3)	ATAY	GTAY	YTAY	
	12	13	14	
5	His	His	His	His
His	Pro	Thr	Ala	Ser
(4)	CCAY	ACAY	GCAY	TCAY
	15	16	17	18
6	Asn	Asn	Asn	
Asn	Gln	Lys	Glu	
(3)	CAAY	AAAY	GAAY	
	19	20	21	
7	Asp	Asp		
Asp	Gly	Arg		
(2)	GGAY	ZGAY		
	22	23		
8	Cys	Cys	Cys	
Cys	Met	Val	Leu	
(3)	ATGY	GTGY	YTGY	
	24	25	26	
9	Gln	Gln	Gln	Gln
Gln	Pro	Thr	Ala	Ser
(4)	CCAX	ACAX	GCAX	TCAX
	27	28	29	30

10	Lys	Lys	Lys				
Lys	Gln	Lys	Glu				
(3)	CAAX	AAAX	GAAX				
	31	32	33				
11	Glu	Glu					
Glu	Gly	Arg					
(2)	GGAX	ZGAX					
	34	35					
12	Ile	Ile	Ile	Ile			
Ile	Tyr	His	Asn	Asp			
(4)	TATM	CATM	AATM	GATM			
	36	37	38	39			
13	Val	Val	Val	Val			
Val	Cys	Gly	Ser	Arg			
(4)	TGTN	GGTN	AGTN	CGTN			
	40	41	42	43			
14	Pro	Pro	Pro	Pro			
Pro	Pro	Thr	Ala	Ser			
(4)	CCCN	ACCN	GCCN	TCCN			
	44	45	46	47			
15	Thr	Thr	Thr	Thr			
Thr	Tyr	His	Asn	Asp			
(4)	TACN	CACN	AACN	GACN			
	48	49	50	51			
16	Ala	Ala	Ala	Ala			
Ala	Cys	Gly	Ser	Arg			
(4)	TGCN	GGCN	AGCN	CGCN			
	52	53	54	55			
17	Gly	Gly	Gly				
Gly	Trp	Gly	Arg				
(3)	TGGN	GGGN	ZGGN				
	56	57	58				
18	Ser	Ser	Ser	Ser	Ser	Ser	Ser
Ser	Phe	Gln	Lys	Glu	Ile	Val	Leu
(7)	TTCN	CAGY	AAGY	GAGY	ATCN	GTCN	CTCN
	59	60	61	62	63	64	65

19	Leu	Leu	Leu	Leu
Leu	Phe	Ile	Val	Pro
(8)	TTTX	ATTX	GTTX	CCTN
	66	67	68	69
	Leu	Leu	Leu	Leu
	Thr	Ala	Ser	Leu
	ACTN	GCTN	TCTN	CTTX
	70	71	72	73

20	Arg	Arg	Arg	Arg	Arg	Arg	Arg
Arg	Gln	Lys	Glu	Pro	Thr	Ala	Ser
(7)	CAGX	AAGX	GAGX	CCGN	ACGN	GCGN	TCGN
	74	75	76	77	78	79	80

W_2

1	Met	Met	
Met	Trp	Cys	
(2)	ATGG	ATGY	
	1	2	
2	Trp		
Trp	Gly		
(1)	TGGN		
	3		
3	Phe	Phe	Phe
Phe	Phe	Ser	Leu
(3)	TTTY	TTCN	TTTX
	4	5	6
4	Tyr	Tyr	Tyr
Tyr	Met	Ile	Thr
(3)	TATG	TATM	TACN
	7	8	9
5	His	His	His
His	Met	Ile	Thr
(3)	CATG	CATM	CACN
	10	11	12
6	Asn	Asn	Asn
Asn	Met	Ile	Thr
(3)	AATG	AATM	AACN
	13	14	15

7	Asp	Asp	Asp			
Asp	Met	Ile	Thr			
(3)	GATG	GATM	GACN			
	16	17	18			
8	Cys	Cys				
Cys	Val	Ala				
(2)	TGTN	TGCN				
	19	20				
9	Gln	Gln	Gln	Gln		
Gln	Asn	Lys	Ser	Arg		
(4)	CAAY	CAAX	CAGY	CAGX		
	21	22	23	24		
10	Lys	Lys	Lys	Lys		
Lys	Asn	Lys	Ser	Arg		
(4)	AAAY	AAAX	AAGY	AAGX		
	25	26	27	28		
11	Glu	Glu	Glu	Glu		
Glu	Asn	Lys	Ser	Arg		
(4)	GAAY	GAAX	GAGY	GAGX		
	29	30	31	32		
12	Ile	Ile	Ile	Ile		
Ile	Phe	Tyr	Ser	Leu		
(4)	ATTY	ATAY	ATCN	ATTX		
	33	34	35	36		
13	Val	Val	Val	Val	Val	Val
Val	Trp	Phe	Tyr	Cys	Ser	Leu
(6)	GTGG	GTTY	GTAY	GTGY	GTCN	GTTX
	37	38	39	40	41	42
14	Pro	Pro	Pro	Pro	Pro	
Pro	His	Gln	Pro	Leu	Arg	
(5)	CCAY	CCAX	CCCN	CCTN	CCGN	
	43	44	45	46	47	
15	Thr	Thr	Thr	Thr	Thr	
Thr	His	Gln	Pro	Leu	Arg	
(5)	ACAY	ACAX	ACCN	ACTN	ACGN	
	48	49	50	51	52	

16	Ala	Ala	Ala	Ala	Ala	
Ala	His	Gln	Pro	Leu	Arg	
(5)	GCAY	GCAX	GCCN	GCTN	GCGN	
	53	54	55	56	57	
17	Gly	Gly	Gly	Gly	Gly	
Gly	Asp	Glu	Val	Ala	Gly	
(5)	GGAY	GGAX	GGTN	GGCN	GGGN	
	58	59	60	61	62	
18	Ser	Ser	Ser	Ser		
Ser	His	Gln	Val	Pro		
(7)	TCAY	TCAX	AGTN	TCCN		
	63	64	65	66		
	Ser	Ser	Ser			
	Ala	Leu	Arg			
	AGCN	TCTN	TCGN			
	67	68	69			
19	Leu	Leu	Leu	Leu	Leu	Leu
Leu	Trp	Phe	Tyr	Cys	Ser	Leu
(6)	YTGG	CTTY	YTAY	YTGY	CTCN	CTTX
	70	71	72	73	74	75
20	Arg	Arg	Arg	Arg	Arg	
Arg	Asp	Glu	Val	Ala	Gly	
(5)	ZGAY	ZGAX	CGTN	CGCN	ZGGN	
	76	77	78	79	80	

W_3

1	Met	Met	2	Trp	Trp
Met	ATG	ATG	Trp	TGG	TGG
(2)	GTA	MTA	(2)	YAC	XAC
	Met	Ile		His	Gln
	1	2		3	4

3	Phe	Phe	4	Tyr	
Phe	TTY	TTY	Tyr	TAY	
(2)	YAA	XAA	(1)	YAT	
	Asn	Lys		Tyr	
	5	6		7	

5	His	His		6	Asn	Asn	
His	CAY	CAY		Asn	AAY	AAY	
(2)	GGT	YGT		(2)	YTT	XTT	
	Trp	Cys			Phe	Leu	
	8	9			10	11	

7	Asp		8	Cys	Cys		
Asp	GAY		Cys	TGY	TGY		
(1)	NCT		(2)	YAC	XAC		
	Ser			His	Gln		
	12			13	14		

9	Gln	Gln		10	Lys	Lys	
Gln	CAX	CAX		Lys	AAX	AAX	
(2)	GGT	YGT		(2)	YTT	XTT	
	Trp	Cys			Phe	Leu	
	15	16			17	18	

11	Glu		12	Ile	Ile		
Glu	GAX		Ile	ATM	ATM		
(1)	NCT		(2)	GTA	MTA		
	Ser			Met	Ile		
	19			20	21		

13	Val		14	Pro		15	Thr
Val	GTN		Pro	CCN		Thr	ACN
(1)	NCA		(1)	NGG		(1)	NTG
	Thr			Gly			Val
	22			23			24

16	Ala		17	Gly			
Ala	GCN		Gly	GGN			
(1)	NCG		(1)	NCC			
	Ala			Pro			
	25			26			

18	Ser	Ser	Ser				
Ser	TCN	TCN	AGY				
(3)	YAG	XAG	NTC				
	Asp	Glu	Leu				
	27	28	29				

19	Leu	Leu	Leu	Leu
Leu	CTN	CTN	TTX	TTX
(4)	YGA	XGA	YAA	XAA
	Ser	Arg	Asn	Lys
	30	31	32	33

20	Arg	Arg
Arg	CGN	AGX
(2)	NGC	NTC
	Arg	Leu
	34	35

W_4

1	Met		2	Trp		
Met	ATG		Trp	TGG		
(1)	TAC		(1)	ACC		
	His			Pro		
	1			2		
3	Phe	Phe		4	Tyr	Tyr
Phe	TTT	TTC		Tyr	TAT	TAC
(2)	AAA	AAG		(2)	ATA	ATG
	Lys	Glu			Ile	Val
	3	4			5	6
5	His	His		6	Asn	Asn
His	CAC	CAT		Asn	AAT	AAC
(2)	GTG	GTA		(2)	TTA	TTG
	Val	Met			Ile	Val
	7	8			9	10
7	Asp	Asp		8	Cys	Cys
Asp	GAT	GAC		Cys	TGT	TGC
(2)	CTA	CTG		(2)	ACA	ACG
	Ile	Val			Thr	Ala
	11	12			13	14
9	Gln		10	Lys	Lys	
Gln	CAX		Lys	AAA	AAG	
(1)	GTY		(2)	TTT	TTC	
	Leu			Phe	Leu	
	15			16	17	

11	Glu	Glu		12	Ile	Ile	Ile
Glu	GAA	GAG		Ile	ATA	ATC	ATT
(2)	CTT	CTC		(3)	TAT	TAG	TAA
	Phe	Leu			Tyr	Asp	Asn
	18	19			20	21	22

13	Val	Val	Val	Val
Val	GTT	GTC	GTA	GTG
(4)	CAA	CAG	CAT	CAC
	Asn	Asp	Tyr	His
	23	24	25	26

14	Pro	Pro	Pro	
Pro	CCC	CCA	CCI	
(3)	GGG	GGT	GGZ	
	Gly	Trp	Arg	
	27	28	29	

15	Thr	Thr	Thr	Thr
Thr	ACT	ACC	ACA	ACG
(4)	TGA	TGG	TGT	TGC
	Ser	Gly	Cys	Arg
	30	31	32	33

16	Ala	Ala	Ala	Ala
Ala	GCT	GCC	GCA	GCG
(4)	CGA	CGG	CGT	CGC
	Ser	Gly	Cys	Arg
	34	35	36	37

17	Gly	Gly	Gly	Gly
Gly	GGT	GGC	GGA	GGG
(4)	CCA	CCG	CCT	CCC
	Thr	Ala	Ser	Pro
	38	39	40	41

18	Ser	Ser	Ser	Ser
Ser	TCI	TCC	AGT	AGC
(4)	AGZ	AGG	TCA	TCG
	Arg	Gly	Thr	Ala
	42	43	44	45

19	Leu	Leu	Leu
Leu	CTT	CTC	YTG
(3)	GAA	GAG	XAC
	Lys	Glu	Gln
	46	47	48

20	Arg	Arg	Arg	Arg
Arg	CGT	CGC	ZGA	ZGG
(4)	GCA	GCG	ICT	ICC
	Thr	Ala	Ser	Pro
	49	50	51	52

W_5

1	Met	Met	Met	Met
Met	ATG	ATG	ATG	ATG
(4)	ACC	ACA	ACG	ACT
	Pro	Thr	Ala	Ser
	1	2	3	4

2	Trp	Trp	Trp	Trp
Trp	TGG	TGG	TGG	TGG
(4)	CCC	CCA	CCG	CCT
	Pro	Thr	Ala	Ser
	5	6	7	8

3	Phe	Phe	Phe	Phe	Phe
Phe	TTT	TTT	TTT	TTC	TTC
(5)	AAC	AAA	AAG	AGG	AGZ
	Gln	Lys	Glu	Gly	Arg
	9	10	11	12	13

4	Tyr	Tyr	Tyr	Tyr
Tyr	TAT	TAT	TAT	TAT
(8)	TAT	TAC	TAA	TAG
	Tyr	His	Asn	Asp
	14	15	16	17
	Tyr	Tyr	Tyr	Tyr
	TAC	TAC	TAC	TAC
	TGT	TGG	TGA	TGC
	Cys	Gly	Ser	Arg
	18	19	20	21

5	His	His	His	His
His	CAT	CAT	CAT	CAT
(8)	TAT	TAC	TAA	TAG
	Tyr	His	Asn	Asp
	22	23	24	25

	His	His	His	His
	CAC	CAC	CAC	CAC
	TGT	TGG	TGA	TGC
	Cys	Gly	Ser	Arg
	26	27	28	29
6	Asn	Asn	Asn	Asn
Asn	AAT	AAT	AAT	AAT
(8)	TAT	TAC	TAA	TAG
	Tyr	His	Asn	Asp
	30	31	32	33
	Asn	Asn	Asn	Asn
	AAC	AAC	AAC	AAC
	TGT	TGG	TGA	TGC
	Cys	Gly	Ser	Arg
	34	35	36	37
7	Asp	Asp	Asp	Asp
Asp	GAT	GAT	GAT	GAT
(8)	TAT	TAC	TAA	TAG
	Tyr	His	Asn	Asp
	38	39	40	41
	Asp	Asp	Asp	Asp
	GAC	GAC	GAC	GAC
	TGT	TGG	TGA	TGC
	Cys	Gly	Ser	Arg
	42	43	44	45
8	Cys	Cys	Cys	Cys
Cys	TGT	TGT	TGT	TGT
(8)	CAT	CAC	CAG	CAA
	Tyr	His	Asp	Asn
	46	47	48	49
	Cys	Cys	Cys	Cys
	TGC	TGC	TGC	TGC
	CGT	CGG	CGC	CGA
	Cys	Gly	Arg	Ser
	50	51	52	53
9	Gln	Gln	Gln	Gln
Gln	CAA	CAA	CAA	CAA
(8)	TTT	TTA	TTG	TTC
	Phe	Ile	Val	Leu
	54	55	56	57

	Gln	Gln	Gln	Gln
	CAG	CAG	CAG	CAG
	TCC	TCA	TCG	TCT
	Pro	Thr	Ala	Ser
	58	59	60	61
10	Lys	Lys	Lys	Lys
Lys	AAA	AAA	AAA	AAA
(8)	TTT	TTA	TTG	TTC
	Phe	Ile	Val	Leu
	62	63	64	65
	Lys	Lys	Lys	Lys
	AAG	AAG	AAG	AAG
	TCC	TCA	TCG	TCT
	Pro	Thr	Ala	Ser
	66	67	68	69
11	Glu	Glu	Glu	Glu
Glu	GAA	GAA	GAA	GAA
(8)	TTT	TTA	TTG	TTC
	Phe	Ile	Val	Leu
	70	71	72	73
	Glu	Glu	Glu	Glu
	GAG	GAG	GAG	GAG
	TCC	TCA	TCG	TCT
	Pro	Thr	Ala	Ser
	74	75	76	77
12	Ile	Ile	Ile	
Ile	ATA	ATA	ATA	
(8)	ATA	ATG	ATY	
	Ile	Val	Leu	
	78	79	80	
	Ile	Ile	Ile	
	ATT	ATT	ATT	
	AAC	AAA	AAG	
	Gln	Lys	Glu	
	81	82	83	
	Ile	Ile		
	ATC	ATC		
	AGG	AGZ		
	Gly	Arg		
	84	85		

13	Val	Val	Val	
Val	GTT	GTT	GTT	
(12)	AAC	AAA	AAG	
	Gln	Lys	Glu	
	86	87	88	
	Val	Val		
	GTC	GTC		
	AGG	AGZ		
	Gly	Arg		
	89	90		
	Val	Val	Val	
	GTA	GTA	GTG	
	ATA	ATG	ATY	
	Ile	Val	Leu	
	91	92	93	
	Val	Val	Val	Val
	GTG	GTG	GTG	GTG
	ACC	ACA	ACG	ACT
	Pro	Thr	Ala	Ser
	94	95	96	97
14	Pro	Pro	Pro	
Pro	CCT	CCT	CCT	
(13)	GAC	GAA	GAG	
	Gln	Lys	Glu	
	98	99	100	
	Pro	Pro	Pro	
	CCC	CCC	CCC	
	GGT	GGG	GGZ	
	Trp	Gly	Arg	
	101	102	103	
	Pro	Pro	Pro	
	CCA	CCA	CCA	
	GTA	GTG	GTY	
	Met	Val	Leu	
	104	105	106	
	Pro	Pro	Pro	Pro
	CCG	CCG	CCG	CCG
	GCC	GCA	GCG	GCT
	Pro	Thr	Ala	Ser
	107	108	109	110

15	Thr	Thr	Thr	
Thr	ACT	ACT	ACT	
(13)	GAC	GAA	GAG	
	Gln	Lys	Glu	
	111	112	113	
	Thr	Thr	Thr	
	ACC	ACC	ACC	
	GGT	GGG	GGZ	
	Trp	Gly	Arg	
	114	115	116	
	Thr	Thr	Thr	
	ACA	ACA	ACA	
	GTA	GTG	GTY	
	Met	Val	Leu	
	117	118	119	
	Thr	Thr	Thr	Thr
	ACG	ACG	ACG	ACG
	GCC	GCA	GCG	GCT
	Pro	Thr	Ala	Ser
	120	121	122	123
16	Ala	Ala	Ala	
Ala	GCT	GCT	GCT	
(13)	GAC	GAA	GAG	
	Gln	Lys	Glu	
	124	125	126	
	Ala	Ala	Ala	
	GCC	GCC	GCC	
	GGT	GGG	GGZ	
	Trp	Gly	Arg	
	127	128	129	
	Ala	Ala	Ala	
	GCA	GCA	GCA	
	GTA	GTG	GTY	
	Met	Val	Leu	
	130	131	132	
	Ala	Ala	Ala	Ala
	GCG	GCG	GCG	GCG
	GCC	GCA	GCG	GCT
	Pro	Thr	Ala	Ser
	133	134	135	136

17	Gly	Gly	Gly	Gly	
Gly	GGT	GGT	GGT	GGT	
(15)	CAT	CAC	CAA	CAG	
	Tyr	His	Asn	Asp	
	137	138	139	140	
	Gly	Gly	Gly	Gly	Gly
	GGC	GGC	GGC	GGC	GGG
	CGT	CGG	CGC	CGA	CCT
	Cys	Gly	Arg	Ser	Ser
	141	142	143	144.1	144.2
	Gly	Gly	Gly	Gly	
	GGA	GGA	GGA	GGA	
	CTT	CTA	CTG	CTC	
	Phe	Ile	Val	Leu	
	145	146	147	148	
	Gly	Gly	Gly		
	GGG	GGG	GGG		
	CCC	CCA	CCG		
	Pro	Thr	Ala		
	149	150	151		

18	Ser	Ser	Ser	Ser	Ser
Ser	TCT	TCT	TCT	TCC	TCC
(18)	GAC	GAA	GAG	GGT	GGG
	Gln	Lys	Glu	Trp	Gly
	152	153	154	155	156.1
	Ser	Ser	Ser	Ser	Ser
	AGC	TCC	AGC	TCA	TCA
	CGG	GGZ	CGC	GTA	GTG
	Gly	Arg	Arg	Met	Val
	156.2	157.1	157.2	158	159
	Ser	Ser	Ser	Ser	Ser
	TCA	TCG	TCG	TCG	TCG
	GTY	GCC	GCA	GCG	GCT
	Leu	Pro	Thr	Ala	Ser
	160	161	162	163	164.1
	Ser	Ser	Ser	Ser	Ser
	AGC	AGT	AGT	AGT	AGT
	CGA	CAT	CAC	CAA	CAG
	Ser	Tyr	His	Asn	Asp
	164.2	165	166	167	168

	Ser AGC CGT Cys 169				
19 Leu (12)	Leu CTT AAC Gln 170	Leu CTT AAA Lys 171	Leu CTT AAG Glu 172		
	Leu CTC AGG Gly 173	Leu CTC AGZ Arg 174	Leu YTA ATA Ile 175	Leu YTA ATG Val 176	Leu YTA ATY Leu 177
	Leu YTG ACC Pro 178	Leu YTG ACA Thr 179	Leu YTG ACG Ala 180	Leu YTG ACT Ser 181	
20 Arg (15)	Arg CGT CAT Tyr 182	Arg CGT CAC His 183	Arg CGT CAA Asn 184	Arg CGT CAG Asp 185	
	Arg CGC CGT Cys 186	Arg CGC CGG Gly 187	Arg CGC CGA Ser 188.1	Arg ZGG CCT Ser 188.2	Arg CGC CGC Arg 189
	Arg ZGA CTT Phe 190	Arg ZGA CTA Ile 191	Arg ZGA CTG Val 192	Arg ZGA CTC Leu 193	
	Arg ZGG CCC Pro 194	Arg ZGG CCA Thr 195	Arg ZGG CCG Ala 196		

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Met	1	3	1		2	24	2	2	1				3		5	5	5		5		
Trp	2	2				3				3				2	45	5	5	1	5	2	
Phe	3			12			3			5	39	45	2	2				5	1	12	5
Tyr	4	1			35	5	5	5	5				16	24		1		5	5	2	5
His	5	14	3		5	5	5	5	35				1	4	2	12	2	5	25		5
Asn	6	1		3	5	5	5	5	5	2	2	2	14	4		1		5	5	3	5
Asp	7	1			5	5	5	5	5				14	4		1		25	35		25
Cys	8	2			5	35	5	5	5	3				12		4	14	5	5	2	5
Gln	9		3	5			1		3		1		5	5	25	25	25		17	45	1
Lys	10			39			1			2	12	2	5	5	5	5	5		15	39	1
Glu	11			45			1				1		5	5	5	5	5	2	18	45	12
Ile	12	3		1	16	2	24	24		5	5	5	35	5				5	1	15	5
Val	13		1	1	14	4	4	4	12	5	5	5	5	5	5	35	5	25	17	15	25
Pro	14	5	45			1				15	5	5		5	17	25	25	39	25	15	19
Thr	15	5	5		2	12	2	2	4	15	5	5		35	15	5	5	45	45	15	19
Ala	16	5	5			1			24	15	5	5		5	15	5	35	29	29	15	75
Gly	17		2	5	5	5	5	15	5			1	5	15	39	45	45	17	45	5	25
Ser	18	5	5	2	5	15	5	35	5	17	25	28	2	17	15	45	19	45	5	65	19
Leu	19		1	12	1		3		1	45	39	45	25	25	25	25	25	5	65	17	35
Arg	20			5	5	5	5	15	5	2	2	12	5	15	29	29	75	15	29	35	35

Рис. 2
Представление 443 э.п., которые содержатся в множествах W_1-W_5.