МАТЕМАТИЧЕСКАЯ БИОЛОГИЯ И БИОИНФОРМАТИКА, 2006, том 1, №1, с.70-96,
http://www.matbio.org/downloads/Kozlov2006(1_70).pdf

==============================БИОИНФОРМАТИКА==============================

УДК 577.21

Математический анализ генетических кодов

©2006 Козлов Н.Н.

ИПМ им. М.В.Келдыша РАН

Обзор завершенного цикла исследований по математическому анализу взаимосвязи структуры генетического кода и необычных способов записи генетической информации - так называемых перекрывающихся генов, когда один и тот же участок ДНК кодирует две белковые последовательности.
Основой для исследований является введенное в рассмотрение множество элементарных перекрытий или перекрытий, соответствующих одиночным аминокислотам. На основе множества доказана теорема, устанавливающая потенциал генетического кода, который использует природа для построения каждого из 5-ти способов перекрывания генов, разрешенных структурой ДНК. Все эти способы обнаружены в экспериментах. На основе теоремы устанавливается непроизвольность структуры стандартного генетического кода. Вычислен аналогичный потенциал для всех нестандартных кодов (по данным Internet таких кодов уже 14). Анализ полученных потенциалов позволил установить одно общее свойство всех известных 15-и природных кодов. Проведен анализ геномов, содержащих парные генетические перекрытия и записанных нестандартными кодами. Показывается неслучайность переосмысленных кодонов или кодонов, отклоняющих код от его стандартной структуры. Устанавливается связь между компактностью геномов некоторых органелл с нестандартностью их генетического кода.

Ключевые слова: *Перекрывающиеся гены, генетический код, вырожденность кода, девиантные коды, происхождение кода, эволюция кода*

Работа выполнена при финансовой поддержке РФФИ (код проекта: 04-01-00320), а также поддержке Программы фундаментальных исследований Президиума РАН «Параллельные вычисления на многопроцессорных вычислительных системах».

Введение. В 1987 году академик Т.М. Энеев предложил мне обратиться к задачам молекулярной биологии. Основная идея состояла в том, чтобы попытаться приложить к этой стремительно развивающейся области науки тот методический материал, который был наработан нами ранее при исследовании некоторых задач по изучению эволюции и структуры сложных природных и технических дискретных систем с большим числом взаимодействующих элементов. Это были задачи о гравитационном взаимодействии галактик [1, 2], по исследованию одной модели процесса аккумуляции планетных систем [3, 4] и по проектированию многослойных интегральных схем [5]. Основой перехода к биологической проблематике должен был стать один метод структурного моделирование эволюции сложных дискретных систем с большим числом взаимодействующих элементов [6, 7]. Впервые высокая эффективность этого метода была установлена на указанной выше модели формирования планетных систем с числом прототел до 10^6 (и это на БЭСМ- 6 с ее оперативной памятью 32 000 слов!). Первая задача по моделированию структуры биологических молекул – моделирование процессов формирования вторичной структуры молекул рибонуклеиновых кислот - РНК была поставлена для коротких молекул – тРНК. Результаты этого исследования вызвали интерес среди специалистов и их признание [8]. В настоящее время нами продолжается исследование усовершенствованных моделей структуризации для современных биохимических данных и более длинных молекул РНК. Это РНК различных типов с размерами, на 1- 2 порядка превышающих тРНК, (5S РНК, молекул-ферментов – рибонуклеаза Р- РНК, 16S РНК). Расчеты таких моделей проводятся на супер-ЭВМ МВС- 1000 [9- 11].

Данное исследование возникло непосредственно в ходе математического анализа необычных способов записи структурных генов (генов, кодирующих белки), и ведет свое начало с работы [12]. Главная цель данной обзорной статьи состоит в том, чтобы изложить один подход к изучению структуры генетического кода. Такой подход оказался весьма плодотворным и в настоящее время привел к важному расширению. Автор не делает попытки исчерпывающего обзора других подходов.

Генетический код. История открытия генетического кода достаточно подробно описана М. Ичасом [13, 14]- одним из участников пионерских исследований по этой проблеме. Он пишет: «…расшифровка биологического кода действительно революционизирующее событие, ее, быть может, уместно сравнить с другим событием, вызвавшим переворот в науке сто лет назад с появлением дарвиновского «Происхождения видов» [13].

Самым трудным в проблеме кода было понять, что код существует. На это потребовалось почти целое столетие. Отсчет его ведется от работы Менделя [15], который показал, что наследственные признаки передаются дискретными частицами, которые мы сегодня называем генами. Эта работа, как известно, почти не вызвала интереса. «Из всего того, что нам известно, складывается впечатление, что Менделю были в общем-то, безразличны отклики на его работу. Опубликовав свой главный труд, он посчитал свой долг исполненным: если на нее не обратили внимания, то тем хуже для читателей, а не для автора». [14, стр. 142]. В 1900 году три независимых исследователя одновременно своими опытами подтвердили результаты, полученные Менделем. Только завершив работу, они узнали, что 34 года назад их опередил Мендель. После 1900 года генетика стала развиваться быстро и непрерывно.

Впервые идея молекулярно-биологического подхода к проблемам генетики была сформулирована известным физиком Э. Шредингером в книге «Что такое жизнь? С точки зрения физика» [16], которая увидела свет в 1945 году. На странице 28 читаем представление о коде (за 21 год до его окончательной разгадки!): «Называя структуру хромосомных нитей шифровальным кодом, мы подразумеваем, что всеохватывающий ум, вроде такого, который некогда представлял себе Лаплас и которому каждая причинная связь непосредственно открыта, мог бы, исходя из структуры хромосом, сказать, разовьется ли яйцо при благоприятных условиях в черного петуха или в крапчатую курицу, в муху или растение маиса, в рододендрон, жука, мышь или человека». Помимо этого и других блистательных предвидений следует отметить, что эта книга сыграла решающую роль в судьбе ряда физиков-теоретиков. Назову лишь две фамилии, о которых будет идти речь в дальнейшем. Это Ф. Крик, который в 1946 году оставил теоретическую физику и обратился к задачам биологии после прочтения этой книги. Его Нобелевская лекция была посвящена проблеме кода, а не структуре ДНК, за которую он был удостоен премии (F.Crick - Nobel Lecture, Dec. 11, 1962: On the Genetic Code, Internet). У истоков проблемы кода стоял также физик Г. Гамов, на которого Ф.Крик ссылается на первой странице указанной лекции. В предисловии автора [13] также читаем: «Вопрос о кодировании стали рассматривать как конкретную проблему, над которой можно работать с надеждой на определенный успех, после заметки Гамова, опубликованной в журнале «Nature» в 1954 г.».

Но сначала была решена проблема структуры ДНК. Аспиранту Д. Уотсону понадобилось всего полтора года, чтобы совместно с руководителем Ф. Криком решить одну из важнейших проблем биологии, которая в настоящее время считается одной из главных фундаментальных проблем, решенных в прошлом столетии. Речь идет о структуре молекул ДНК, которую мир впервые увидел 25 апреля 1953 года: работа [17], объемом в одну (!) страницу журнала «Nature» поставила точку на дискуссии относительно роли ДНК в передаче наследственной информации. Точка была поставлена в двадцатипятилетнем споре относительно ее структуры, когда, как считают современные биологи, благодаря неверной гипотезе 1931 года, было задержано развитие молекулярной биологии на целую четверть века [13]. Сама же ДНК (дезоксирибонуклеиновая кислота), одна из двух (еще и РНК) нуклеиновых кислот, была открыта в 1868 году.

Описания, которые дают для ДНК сегодня, различны. Для наших целей достаточно упрощенного описания. Модель двойной спирали ДНК представляет собой две нити, закрученные друг относительно друга (рис. 1). По сути дела, это – двойная винтовая линия, а не какая ни спираль. Алфавит ДНК содержит всего 4 буквы: A, C, G, T. Это четыре нуклеотида: аденин, цитазин, гуанин, и тимин. Точки между этими буквами на рис. 1 указывают на количество водородных связей: две связи между A и T и три между C и G. Именно эта блестящая догадка Уотсона, который ввел эти комплиментарные пары [18], и позволила объяснить важнейшие свойства передачи наследственной информации. (Эти связи существуют между двумя спиралями ДНК). ДНК измеряют по разному, в том числе, и количеством пар нуклеотидов. Например, для ДНК человека их около 3.2 миллиардов (ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/).

Тайна гена была окончательно разгадана в 1966 году (к столетию работы Менделя [15]), когда в ходе экспериментальных исследований было окончательно установлено, что гены есть однонитиевые участки ДНК и содержат информацию о белке в закодированном виде. Оказалось, что каждая из 20 аминокислот - элементов, из которых состоят все известные белки, кодируется определенными тройками нуклеотидов - кодонами или триплетами. Для четырех букв: A, C, G, T, имеем 64 кодона: AAA, AAC, AAG, … TTT. Смысл всех этих кодонов был экспериментально установлен и представлен в таблице генетического кода, причем кодировка, которую выбрала природа оказалась достаточно своеобразной. В табл. 1 в столбце K^0 она представлена полностью. Оказалось, что только две аминокислоты - метионин (Met) и триптофан (Trp) кодируются однозначно кодонами ATG и TGG соответственно. Все остальные аминокислоты кодируются более чем одним кодоном (это кодоны-синонимы), но не более чем шестью. Последнее наблюдается только для трех аминокислот: серин (Ser), лейцин (Leu), аргинин (Arg). Такие три кодировки названы нерегулярными, в отличие от 17-и других, регулярных для которых каждые 1-ая и 2-ая позиции одинаковы в соответствующем наборе кодонов-синонимов. Полное число смысловых кодонов (т.е. троек кодирующих какую-либо аминокислоту) равно 61, а три кодона TGA, TAA, TAG не кодируют никакую из аминокислот, а присутствуют в конце генетического текста и указывают признак конца при белковом синтезе (белок синтезируется на основе текста гена по правилу генетического кода до тех пор, пока в тексте гена не возникнет одна из названных троек). Эти три тройки получили название терминаторных кодонов.

Укажем, что помимо вырожденности (т.е. когда одной и той же аминокислоте соответствуют, как правило, несколько кодонов-синонимов) важнейшим свойством кода является его универсальность: код одинаков для почти всех живых организмов. Однако к настоящему времени обнаружены ряд отклонений кода от стандартного - K^0, что является одной из наиболее загадочных особенностей кода. Три подобных нестандартных кода K^1-K^3 (K^1 закодированы белки в митохондрии человека) приведены также в табл. 1. Некоторый анализ, основанный на этих кодах приводится ниже.

На рис.2 представлен начальный участок одного гена и по правилу генетического кода выписан участок соответствующей белковой последовательности. Начальная позиция (жирная точка на рис.), откуда начинается белок, устанавливается экспериментально.

Перекрывающиеся гены.Проведенный математический анализ структуры генетического кода основывался на исследовании некоторых необычных способов записи структурных генов. Необычный способ записи генов состоит в том, что один и тот же участок цепи ДНК, кодирующий белок, может читаться со сдвигом фазы на +1 либо -1 нуклеотид либо с другой цепи ДНК (с подобными сдвигами либо без них). Иными словами один и тот же указанный участок ДНК может кодировать два и более негомологичных белка - вплоть до шести. 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, который происходит (по тексту гена). Иными словами белок при альтернативном чтении не синтезируется. Был сделан вывод о мощной биологической защите: природе не нужны эфемерные белки, она не синтезирует белки соответствующие сдвинутым позициям (например, если в ходе мутаций начальная точка сдвинется). Тем самым были введены в рассмотрение последовательности кодонов или две рамки считывания - РС: открытая рамка считывания (ОРС) - последовательность кодонов, не содержащая кодонов терминации и блокированная РС - БРС, когда такие кодоны встречаются [19]. На рис. 2 белок соответствует ОРС, сдвинутые позиции как +1 так и -1 - БРС. Оказалось, что лишь для перекрывающихся генов такого запрета не существует. Впервые этот эффект был установлен в 1976 году в ходе исследований по чтению первого целого генома - вируса бактерии ФХ 174 [20]. ДНК такого вируса оказалась кольцевой и одноцепочечной и содержащей 5386 нуклеотидов [21]. Руководитель этих исследований F.Sanger был удостоен второй Нобелевской премии по той же специальности что и ранее (по химии). Отмечу, что это один из двух лауреатов, для которых было сделано подобное исключение за всю историю таких премий [22].

На рис. 3 представлено первое из обнаруженных генетических перекрытий: это перекрытие генов D и E в геноме ФХ174. Приводится фрагмент из окончательной публикации всего этого генома [21]. РС соответствующая белку E сдвинута на +1 нуклеотид относительно ОРС для белка D. Начиная с позиции 567 возникает измененная последовательность триплетов, которая не содержит кодонов терминации, т.е. является также ОРС как и набор триплетов для D. Таким образом ген E целиком лежит внутри гена D. Одно из самых длинных перекрытий, обнаруженных к настоящему времени, относится к ДНК вируса GSHV [23]- одного из вирусов группы HBV - вирус гепатита B человека. Этот необычный вирус вызывает болезни печени и распространенную форму рака. Весь геном GSHV содержит 3311 пар нуклеотидов. Перекрытие генов составляет 1704 нуклеотида (для близкородственного вируса ASHV [24] - 1698 нуклеотида), причем наибольшее перекрытие содержит 428 кодонов (около 1300 нуклеотидов): это перекрытие генов S и A; ген S целиком (со сдвигом на +1 нуклеотид) принадлежит гену A и также наблюдается частичное перекрывание генов C и В с геном А. Иными словами если бы перекрытия отсутствовали, то размер генома GSHV должен был бы возрасти более чем в 1,5 раза (на 1704 нуклеотида), т.е. перекрывающиеся гены являются важнейшим фактором сокращения кодирующего объема ДНК. Вопрос о том, как это могло произойти является важнейшей проблемой биологии. Лишь некоторые аспекты этой проблемы анализируются ниже.

О востребованности каждого из 64 кодонов в генетических перекрытиях. Важной работой в данном цикле исследований стала работа [25]. Представим кратко основные позиции этой работы.

Результаты по исследованию генов, принадлежащих одной цепи ДНК и попарно перекрывающихся [12, 26, 27], позволили выявить ряд важных особенностей. Было установлено, что для подобных генов существуют около 300 различающихся локальных перекрываний, каждое из которых содержит не более 4-х кодонных семейств, и однозначно определяет потенциальные позиции (одну либо две), а также тип нуклеотидных замен, соответствующих молчащим мутациям. Такие замены не влияют на одну либо две белковые последовательности в перекрытии. Анализ десятков геномов, содержащих генетические перекрытия, показал, что число таких позиций относительно невелико. В каждой из таких позиций допускается использование пар кодонов - синонимов, в любых других позициях такое использование неизбежно приводит к искажению одной либо двух аминокислотных последовательностей, закодированных в перекрытии. Такая почти жесткая связь между кодонами позволила обратиться к изучению роли каждого из 64 кодонов универсального генетического кода в случае экспериментально обнаруженных перекрываний генов. Причем наибольший интерес представляют перекрытия без указанных позиций. Проведенные расчеты показали, что наиболее протяженная область перекрывания с указанным свойством содержится в геноме RSV [28]: это перекрывание генов pol и env на промежутке в 137 нуклеотидов; в перекрытии использовано более 90 кодонов. В структуре такого перекрывания было использовано 42 (из 61) различающихся смысловых кодона или было использовано вдвое больше минимально возможного числа кодонов (20), необходимого для записи аминокислотной последовательности геном, который не является перекрывающимся.

Были выделены случаи, где число позиций с потенциальными молчащими мутациями является относительно небольшим. Анализ десятков геномов, содержащих перекрытия [29] показал, что таким свойством обладают перекрывания из 814 нуклеотидов в геномах близкородственных фагов ФХ174 [21] и G4 [30]: одна позиция с указанным свойством на 90-100 нуклеотидов или таких позиций втрое меньше чем в случае вируса GSHV. Рассмотрим вопрос о частоте встречаемости каждого из 64 кодонов универсального генетического кода в названных геномах. Этот вопрос рассматривался ранее [30] для полных наборов генов в этих фагах. В табл. 2 приводится частота встречаемости соответствующих кодонов лишь для областей перекрывания генов В, К, Е для ФХ174. Из табл. 2 следует, что в структуре генетических перекрытий ФХ174 использован весь набор смысловых кодонов - 61 кодон. Этот вывод не может считаться окончательным, т.к. не изучены возможности позиций с потенциальными молчащими мутациями, которые в принципе, в силу возможности использования в таких позициях кодонов-синонимов, могут изменить полное число используемых кодонов. Укажем полный набор таких возможностей, который был вычислен на основе указанных правил, по перекрытиям в ФХ174. На рис. 4 представлены все локальные перекрывания, в которых допускается использование кодонов-синонимов в случае ФХ174. Имеем 9 позиций с допустимыми заменами, причем лишь одна из них связана с ter кодонами. В случае ФХ174 могут быть использованы 17 смысловых кодонов-синонимов. Однако, влияние кодонов-синонимов не изменяет принципиальности результата: для записи перекрытий в ФХ174 не может быть использовано число смысловых кодонов меньше их полного набора - 61 кодон (см. подпись под рис. 4). Можно показать также [25], что в случае G4 число используемых смысловых кодонов - 60, не был использован лишь кодон GTA (Val).

Рассмотрим роль тройки терминаторных кодонов TAA, TGA, TAG. На рис.5 представлены локальные перекрытия, из G4 и HeV [31], соответствующие областям терминации. Видим, что все три терминаторных кодона являются не заменяемыми в конструкциях приводимых трех перекрываний.

Полученные результаты свидетельствуют о том, что все 64 кодона универсального генетического кода востребованы в генетических перекрытиях: запись соответствующих генетических перекрытий не представляется возможной при исключении хотя бы одного кодона из 64. Такого жесткого требования к участию каждого из 64 кодонов не выдвигается при изучении неперекрывающихся генов. Таким образом, обнаружена жесткая взаимосвязь между полным набором кодонов и генетическими перекрытиями. Из этих результатов также следует, что без вырожденности кода перекрытия генов становятся допустимыми только на относительно небольших интервалах и построение протяженных генетических перекрытий становится практически невозможным. В связи с этим в работе [25] сделан вывод, что перекрывающиеся гены явились одним из факторов, повлиявших на формирование окончательной структуры генетического кода с его вырожденностью. В ходе дальнейших исследований этот вывод получил свое развитие.

Множества, порождаемые генетическим кодом. Основные результаты данной работы были получены на основе математического анализа множеств кратко представленных ниже. Впервые ограниченные множества подобного типа анализировались ранее [32, 33]. При этом речь шла только о перекрытиях генов, принадлежащих одной цепи ДНК. Таких случаев 2: сдвиг на +1 либо -1 нуклеотид относительно исходных генов. Позднее были установлены перекрытия двух генов, принадлежащих различным цепям ДНК, которые называются плюс-цепью и минус-цепью ДНК. При этом следует отметить, что чтение гена (и последовательностей триплетов) происходит слева направо для плюс-цепи ДНК и справа налево для минус-цепи ДНК (согласно модели из рис. 1). Случаев перекрывания при этом будет 3: сдвиг на +1, 0 либо -1 нуклеотид гена из минус-цепи относительно гена из плюс-цепи. Таким образом, полное число случаев перекрытий пар генов равно 5. Экспериментальные данные по всем таким случаям представлены на рис. 6; это данные лишь по двум геномам: mtDHK Bovin [34] и IS5 [35, 36].

Было введено понятие элементарного перекрытия - э.п. для описания всех 5-и случаев перекрывания, а также все множества 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. 7 дано краткое представление э.п.: представлены лишь по 4 э.п. для каждого из множеств W_1-W_5. Полный набор э.п. оказался равным 448, а 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 - 201. Анализ этих множеств дается в [37] на стр. 13-27 в [37] приводится полный перечень э.п. для этих множеств. Интересно было сравнить численность всех э.п. для одной и двух цепей ДНК. Формальное сравнение это 160 и 288, а фактическое - это 80 и 288 т.е. число э.п. для двух цепей ДНК более чем в 3,5 раза больше. Это связано с тем, что множества W_1 и W_2 фактически одинаковы; их отличает только перестановка аминокислот в строках. 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. На рис. 8 дано сжатое представление для всех 448 э.п. По оси абсцисс указывается номер аминокислоты из верхней строки в э.п. - AA^t, t - top (верхний), по оси ординат - номер аминокислоты из нижней строки в э.п. - AA^l, l - lower (нижний). Число возможных позиций (клеток) равно 400. Ясно, что для 448 э.п. некоторые из позиций должны быть заняты э.п., принадлежащим различным множествам. Такие 113 позиций указаны двухзначными числами. Кроме того 182 позиции заняты числами 1-5, которые соответствуют номерам множеств W_1-W_5. Таким образом, из 400 возможных позиций только 295 (113+182) заняты какими-либо э.п. (они заштрихованы на рис. 8), а оставшиеся 105 позиций являются свободными. Иными словами э.п. из всех множеств W_1-W_5 не содержат все возможные перекрывания (когда заполнены все 400 клеток) любых пар из 20-и 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.

Отметим, что парные генетические перекрытия могут быть исследованы при решении задачи сборки э.п., принадлежащих одному множеству (см. рис. 9). Такая сборка или стыковка э.п. осуществляется при помощи наборов нуклеотидов в 1-х и 4-х позициях среди которых непременно должен быть одинаковый нуклеотид (на рис. 9 э.п. с номером 57 содержит в четвертой позиции N или любой из четырех нуклеотидов, в т.ч. и C, который содержится в первой позиции следующего э.п. - 19). В ходе такой сборки было поставлено два вопроса: каков потенциал генетического кода для построения перекрытий и каковы сочетания аминокислот при этом могут быть в верхней строке перекрытия. Ответ на второй вопрос оказался следующим [33]: любые. Для ответа на первый вопрос было вычислено полное множество перекрытий для 428 кодонов на основе W_1. Расчеты показали [33], что это множество содержит около $\sim 10^{746}$ перекрытий генов (см. подпись к рис.9), каждый из которых имеет протяженность 428 кодонов; одно из таких перекрытий соответствует генам S и A в геноме GSHV [23]. При этом среднее число перекрытий такой протяженности имеющих одинаковые последовательности аминокислот в первой строке (например, одинаковые последовательности S) составляет $\sim 10^{189}$.

Феноменальный потенциал для построения на основе W_1 парных генетических перекрытий в случае сдвига -1 нуклеотид, а также отсутствие какого-либо запрета на возможность такого перекрывания для любых белковых последовательностей, выдвинул задачу исследования всех возможных способов парных генетических перекрытий, допускаемых структурой ДНК.

Интегральная характеристика генетического кода. Проведенное исследование позволило установить одну интегральную характеристику K^0. Тем самым одно число будет определять целую сложную структуру кодонно-аминокислотных соответствий, определяющих K^0, такое число будет изменяться при отклонениях кода от K^0.

Обратимся (см. выше) к понятию рамка считывания РС и двум их типам: открытой - ОРС либо блокированной БРС. Вследствие трехбуквенной кодировки аминокислот, одному и тому же гену соответствуют 6 РС. Для определенности назовем РС0 - РС, соответствующая заданному гену, а РС1 РС5 - 5 альтернативных рамок считывания. При рассмотрении парных генетических перекрытий для каждого из 5 случаев перекрываний (см. рис. 6) имеем по 2 РС: в случае 1 это РС 0 и РС 1, в случае 2 это РС 0 и РС 2, и.т.д., в случае 5 это РС 0 и РС 5.

В генетическом эксперименте по перекрыванию, для каждого РС 0 каждая из альтернативных РС является ОРС, так как они соответствуют реальным белковым последовательностям. Отметим, что из факта существования ОРС вовсе не следует, что она кодирует природный белок; этот вопрос решается в генетическом эксперименте. Важно было выяснить: каков потенциал кода $К^0$ для построения ОРС в альтернативной РС для каждого из пяти случаев перекрываний. Ответ на этот вопрос дает Теорема для генетического кода (впервые в препринте № 80, 2001г., ИПМ им. М.В.Келдыша РАН, копия в[38]); случаи перекрытий в одной цепи ДНК были рассмотрены ранее [33, 39]. Из Теоремы следует, что **код $К^0$ устроен так, что практически для любой последовательности аминокислот, каждая из РС 1- РС 5 при парном перекрывании генов может быть ОРС**. Исключения возникают из- за присутствия в РС 0 хотя бы одной пары аминокислот, которые назовем блокировочными. Эти блокировки возникают только в РС 2, РС 3, и РС 5 и не имеют место в РС 1 и РС 4. Согласно [38] имеем: РС 2 становится БРС для пяти пар:

MetMet, MetAsn, MetLys, MetIle, MetThr, (1)

поскольку в РС 2 образуется кодон ter -TGA; РС 3 становится БРС для 6 пар:

PheTyr, TyrTyr, HisTyr, AsnTyr, AspTyr, CysTyr , (2)

поскольку в РС 3 образуется один из кодонов ter: TAA или TAG; РС 5 становится БРС для 5 пар:

PheMet,PheAsn, PheLys, Phe Ile, PheTyr. (3)

поскольку в РС 5 образуется один из кодонов ter: TAA или TGA. На рис. 10 дано представление всех этих пар с соответствующими кодировками.

Для каждого из генетических кодов введем в рассмотрение числовую интегральную характеристику которую обозначим p, как число различающихся блокировочных пар. Согласно (1)- (3), для $К^0$ имеем p^0= 16. Следует отметить, что уже около 50 лет назад было установлено присутствие всех 400 пар аминокислот в природных белках. Поэтому существование указанных 16 пар аминокислот для $К^0$ должно иметь какое-либо иное объяснение. Важным моментом для этого является анализ целого спектра характеристик p. Ниже анализируются характеристики p для двух важнейших наборов кодов: для гипотетических и природных.

Гипотетические коды. Рассмотрим три набора гипотетических кодов, построенных на основе математического анализа множеств э.п., с целью изменения характеристики p. Теоретически возможный диапазон изменения этой целочисленной характеристики: от 0 до 400.

Из сотен кодов, которые были исследованы нами в табл. 3 приведены лишь 6: 3 пары кодов с различными характеристиками. Первая пара - $К_{11}$, $К_{12}$ - коды, возникающие из $К^0$ с помощью перестановки всего одного кодона; вторая - $К_{21}$, $К_{22}$ - относится к кодам у каждого из которых характеристика p=0 и третья - $К_{31}$, $К_{32}$ - соответствует $p>p^0$ вплоть до почти десятикратного увеличения p (*у* $К_{32}$).

Код $К_{11}$ соответствует перестановке среди смысловых кодонов: TGC(Cys)→Trp, а код $К_{12}$ соответствует перестановкам смыслового кодона в набор ter: CGA(Arg)→ter, т.е. набор ter расширяется до 4-х. Каждый из кодов $К_{11}$-$К_{12}$ соответствует $p>p^0$. Именно такие перестановки и выбирались (Полное число подобных одиночных перестановок для $К^0$ составляет около 140, это число основано на анализе структуры множеств W_1-W_5). В коде $К_{11}$ возникают дополнительно к $К^0$ еще 6 блокировочных пар аминокислот. Максимальное за счет одной перестановки увеличение p дает случай $К_{12}$ (CGA(Arg)→ter): с p^0=16 до 28. При этом ter: YGA, Y: T,C дополнительно блокирует только РС 1 для 12 пар аминокислот.

Обратимся к гипотетическим кодам K_{21}-K_{22} для каждого из которых имеем минимально возможное значение p: p=0. Параметр ν - число кодонных семейств кода, которые отклоняют структуру от регулярной - приводится в скобках в первой строке таблицы. Для K^0 и K_{21} имеем ν=4, для K_{22} ν=0. Рассмотрим K_{21}. Это единственный код с p=0, который может быть образован из K^0 с помощью перестановки всего одного кодона [40]. Можно показать, что для набора $\mathtt{ter}^0$: TAA, TGA, TAG (набор ter для K^0) не существует кода с характеристикой p=0 и ν=1; эта единственная нерегулярность относится к указанному набору ter из K^0.

Наконец, была поставлена задача поиска гипотетических кодов, у которых p много больше p^0. При этом решение искалось не среди всех возможных наборов кодонных семейств, а с учетом структур природных нестандартных кодов из [41]. Для таких кодов набор ter не более 4 кодонов, набор смысловых кодонов для одной аминокислоты не более 8 кодонов. Коду K_{31} соответствует p=58, отличие от K^0 в трех кодонных семействах: Trp, Cys, ter, параметр ν=3 уменьшение с ν^0=4 связано с изменением ter: TGN, N: A, C, T, G. Для K_{32} с тем же набором ter имеем p=155 или почти на порядок больше p^0=16, а $\nu=\nu^0$=4. Такое увеличение p достигнуто за счет изменения 12-и смысловых кодонных семейств в K^0.

Таблица 3 иллюстрирует лишь некоторые из гипотетических кодов, которые были изучены и для которых диапазон изменения характеристики p: от 0 (это min p) до 155 (почти 10-и кратное увеличение по сравнению с p^0). Математический анализ гипотетических кодов весьма важен в связи с непрерывными публикациями новых природных нестандартных генетических кодов. Кроме того, в последние годы были начаты экспериментальные работы по созданию некоторых версии генетического кода, отклоненных от K^0 [42]. В частности, было получено расширение числа кодируемых аминокислот у E.coli и указывается [42], что разработанный подход может быть положен в основу метода расширения генетического репертуара живых клеток и встраивания аминокислот с новыми структурными, химическими и физическими свойствами в белки.

Функциональная роль переосмысленных кодонов. Свойство всех известных природных кодов. Гипотетические коды дают в целом спектр изменения p для различного числа кодонных перестановок. Оказалось, что все известные на сегодня природные генетические коды обладают одним свойством; первый природный нестандартный код был обнаружен у клеточной органеллы человека - митохондрии [43]. Перед представлением этого свойства рассмотрим роль переосмысленных кодонов в перекрываниях генов.

В табл. 1 на основе [43-46] представлены в сравнении с K^0 данные о трех таких кодах: K^1- K^3, причем указаны лишь переосмысленные кодонные наборы: для K^1 их 5, для K^2 и K^3- по 6. Укажем лишь переосмысленные кодоны: для K^1 это ATA(Met), TGA(Trp), AGA и AGG (ter); для K^2 это ATA(Met), TGA(Trp), AGA и AGG(Ser); для K^3 это TGA(Trp), AAA(Asn), AGA и AGG(Ser). Указанные природные перестановки кодонов приводят к изменению размеров кодонных наборов по сравнению с K^0: число кодонов ter равно 2 для K^2, K^3 и 4 для K^1, число кодонов Ser увеличивается до 8 как для K^2, так и для K^3.

Применение способа доказательства Теоремы из [38] для кодов K^1- K^3 позволило получить соответствующие характеристики p. Оказалось, что p^1= p^2= 7, p^3=5, то есть наблюдается уменьшение характеристики p в 2.3- 3.2. раза. Что означает подобное уменьшение блокировочных пар? Анализ показал, что такое уменьшение p приводит к возможности построения для кодов K^1- K^3 перекрытий генов, запрещенных при использовании K^0. На рис. 11 приведены некоторые генетические перекрытия из [44-47], записанные в геномах с использованием указанных нестандартных кодов. Жирным шрифтом отмечены переосмысленные кодоны, а также пары аминокислот: MetAsn (у $К^1$) MetLys (у $К^2$) MetThr (дважды у $К^3$). Эти три пары создавали блокировки в РС 2 при использовании стандартного кода (см. (1) и рис. 10), и стали возможным благодаря одной и той же перестановке TGA(ter) → Trp. Первый фрагмент соответствует Mito Human (впервые [48]). Структуры перекрытий из рис. 11 насыщены переосмысленными кодонами, особенно в случае использования $К^2$ для A.Mellitera ligustica [45] (из 19 нуклеотидов лишь 8 не относятся к переосмысленным кодонам) и $К^3$ для P.lividus [46] (3 переосмысленных кодона участвуют в перекрытии наряду с 3-мя стандартными кодонами, то есть кодонами из $К^0$). При использовании $К^1$ приведены данные для Human [44] и R.Norvegicus [47]. Два фрагмента перекрытий из рис. 11 для R.Norvegicus соответствуют наиболее протяженному перекрыванию – 73 нуклеотида. Первоначальный вывод, который следует из анализа перекрытий из рис. 11 состоит в том, что именно переосмысление кодонов позволяет строить перекрытия для кодов $К^1$-$К^3$, так как они, согласно Теореме, запретны для $К^0$.

Перед окончательным выводом необходимо иметь в виду следующее. Указанные рассуждения справедливы при условии, если в областях, указанных жирным шрифтом, отсутствуют потенциальные позиции молчащих мутаций. Полный набор таких позиций для допустимых перекрытий пар генов из одной цепи ДНК был установлен ранее для $К^0$ [12]. Применение этого подхода для $К^1$- $К^3$ показало, что для этих нестандартных кодов подобные позиции могут также существовать. Например, для перекрытия, записанного $К^1$ из рис. 11 замена T на C в позиции 7916 не изменит ни одну из двух белковых последовательностей, так как эта замена повлияет лишь на смену кодировок в двух аминокислотах: у Ile ATC заменит кодон-синоним ATT, у Leu CTA заменит - TTA. Или другое влияние изменения кодировки: замена T на C в позиции 13529 соответствует позиции молчащей мутации для кода $К^0$ и не имеет место для $К^1$. В связи с этим был проведен анализ перекрытий, записанных $К^1$- $К^3$ и вычислены все позиции молчащих мутаций (для $К^1$ см. [48]). Оказалось, что ни одна из таких потенциальных мутаций не расположена в областях использования переосмысленных кодонов, то есть структура перекрытий в таких областях является «жесткой».

Итак, нами показано (впервые для Mito Human в [48]), что существование рассматриваемых перекрытий невозможно без переосмысления кодонов. В табл. 4 приведены некоторые данные по перекрытиям генов в 6 геномах [44-47, 49-50]: по двум геномам для каждого из кодов $К^1$- $К^3$. Перекрытия генов приводят к сокращению размеров ДНК, на что биологи указали непосредственно после открытия таких генов. Этот эффект оказался более значительным для кодов $К^1$- $К^3$. За пределами рис. 11 остались генетические перекрытия, которые имеют место в некоторых из указанных геномов, и на структуру которых не повлиял фактор переосмысления кодонов. Размер таких перекрытий приводится в скобках в табл. 4. Например, этот размер почти на порядок меньше, чем для перекрытий из рис. 11, записанного кодом $К^1$ (R.Norvegicus, [47]): 8 по сравнению с 73. Относительно величины сокращения размера ДНК следует сказать особо. Все геномы из табл. 4 относятся к так называемым митохондриальным ДНК, а согласно современным данным, число митохондрий в одной- единственной клетке, может быть несколько тысяч [51]. Поэтому эффект указанного сокращения для одной клетки может возрасти на 3- 4 порядка, по сравнению со значениями, указанными в таблице 4. Таким образом, проведенное исследование показало, что переосмысление кодонов у нестандартных кодов не носит случайного характера, как это принимается рядом исследователей [52].

Помимо установленной нами роли переосмысленных кодонов в генетических перекрытиях оказалось, что всех природных генетических кодов (известных к настоящему времени) имеет место одно общее свойство. Для его формулировки отметим, что новизна результатов, полученных при анализе записей генов кодами $К^1$-$К^3$, поставила задачу изучения всех известных из экспериментов кодов.

По Internet данным (http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?mode=t) полное число таких кодов около 20. Для наших исследований были выделены лишь те коды для которых каждый из кодонов является осмысленным, т.е. из исследования были исключены коды в которых участие каких-либо кодонов в кодировках не установлено. Кроме того, отклонение кода от $К^0$ должно состоять в перестановке минимум одного из 64-х кодонов. В итоге осталось всего 14 нестандартных кодов (в них вошли указанные выше 3 кода $К^1$-$К^3$). На основе наших исследований были получены характеристики p для каждого из названных кодов. Расчеты показали, что лишь для одного из них (Thraustochytrium Mito Code): имеем p=21, для всех оставшихся 13 кодов p не превышает p^0=16.

Проведенное исследование показало правильность введения интегральной характеристики генетического кода - p. Имеем для всех 15 известных природных кодов p не превышает 21 или всего не более около 5% пар (21/400) являются запретными по какому-либо одному или редко нескольким из 5-и способов перекрывания генов. Тем самым установлено общее свойство всех 15 природных генетических кодов. Этот результат безусловно имеет фундаментальное значение и в научном мире неизвестен. Анализ полученных характеристик показал, что фактор перекрываемости, возможно, играет помимо прочего какую-то другую важную роль в функционировании генов. Это выдвинуло на передний план новые задачи, которые в настоящее время решаются.

Два фундаментальных вывода. Главные биологические выводы впервые были сформулированы нами в 1999-2000 г. на основе анализа только перекрытий генов, принадлежащих одной цепи ДНК [33, 39, 48]. Дальнейшие исследования, связанные с анализом перекрытий генов, принадлежащих различным цепям ДНК, или в итоге со всеми 5-ю способами перекрывания пар генов, только углубили сформулированные ранее утверждения.

Указанные выводы основываются на анализе двух положений из известной монографии [52]:

1. «The code seem to have been selected arbitrarile…» («Код по-видимому был «выбран» произвольно…»).
2. Переосмысление кодонов «указывают на то, что в генетическом коде митохондрий могут происходить случайные перемены».

Наше исследование обе эти позиции никак не подтверждает и нами были сформулированы два вывода (см. I, II ниже) по существу рассматриваемых вопросов.

I **Генетический код не был «выбран» произвольно**. Такой вывод был сделан первоначально в [33] на основе анализа около 140 одиночных перестановок кодонов в $К^0$, для каждой из которых характеристика $p>p^0$=16 (см. $К_{11}$, $К_{12}$ из табл. 3). Дальнейшие исследования лишь расширили и углубили это утверждение, хотя уже неустойчивое поведение характеристики p даже при одиночных перестановках никак не могло быть в пользу случайного «выбора» $К^0$. На основе завершенного этапа исследований сформулируем следующий вывод. Одним из решающих факторов «выбора» $К^0$ является возможность для кодонных семейств практически беспрепятственно записывать две белковые последовательности одним геном, причем для этого может быть использован наиболее благоприятный (по сочетанию аминокислот в перекрытии) один из 5-и вариантов такой компактной записи генов (5 случаев перекрываний). Категорический запрет существует не более чем для около 5% пар аминокислот как для стандартного, так и для всех 14-и известных на сегодня нестандартных кодов. Т.е. 15 таблиц кодов удовлетворяют одному и тому же общему свойству. Это не оставляет никаких шансов для какой-либо произвольности, случайности «выбора» не только K^0, но и любого из известных девиантных кодов.

Не покидает такое ощущение, что в ходе «выбора» генетического кода была решена (кроме всего прочего) математически довольно сложная задача. Кратко опишем ее отталкиваясь от случайного кода К'. Феноменальное многообразие генетических кодов (число способов разместить 64 кодона в 20 клеток) дает основание предположить, что характеристика p для К' будет $p'>>1$. Если это так, то получение K^0 из К' математически сводится к задаче поиска некоторого локального минимума p (p_{min}=0, а p^0=16 по сравнению с теоретическим максимумом p_{max}=400). Анализ сотен гипотетических кодов (некоторые из которых были выделены в табл. 3) показал, что математический расчет K^0 по К' с целью достижения указанного эффекта по перекрываемости генов сводится к оптимизационной задаче с огромным количеством локальных минимумов. Решение подобных оптимизационных задач даже в существенно более простых случаях весьма затруднительно [53].

Основой нашего вывода явился новый подход к исследованию генетического кода: новые методы биоматематики, разработанные нами (прежде всего теория элементарных перекрытий кратко представленная выше) позволили выйти на уровень, когда на решение наших задач влияет всего один нуклеотид или одна нуклеотидная замена, тем более триплет, кодон-синоним, аминокислота. Такой подход был выработан мною после изучения истории молекулярной генетики. Конкретно имеется ввиду та болезнь, изучение которой способствовало интенсивному развитию молекулярной генетики в последние более чем 60 лет. Это одно из тяжелейших заболеваний человека - серповидноклеточная анемия. Люди, страдающие этой болезнью, как правило погибают, не достигнув зрелого возраста. Окончательная причина заболевания была выяснена только в 1964 году, когда было впервые получено прямое подтверждение того, что гены и белковые последовательности действительно коллинеарны. Было установлено, что замена одного нуклеотида в гене, а именно А на Т, приводит к изменению всего одной аминокислоты Glu на Val, что и приводит к указанному заболеванию (краткое описание со ссылками см. [54]). Это исследование наглядно продемонстрировано, что при работе с генами только выход на изучение роли, влияние всего одного нуклеотида может дать какие-то новые результаты. А перекрывающиеся гены явились весьма удобным объектом исследования, т.к. они дают почти однозначную взаимосвязь между двумя белковыми последовательностями и участком ДНК [12]. В этой области легче использовать математику; в области, где гены не перекрыты, фактор использование в кодировке того или иного кодона-синонима все еще является тайной. Введение в рассмотрение элементарных перекрытий из [37] (см. выше) сыграло решающую роль в данном исследовании.

II **Переосмысление кодонов не является случайным**. Проведен анализ парных генетических перекрытий, записанных с участием нестандартных кодов. Для ряда нестандартных кодов установлена роль этого переосмысления в расширении потенциала для генетической перекрываемости в т. числе показано существование в около двух десятков mtДНК (в т.ч. человека) записей парных генетических перекрытий, которые невозможны (по Теореме [38]) для кода стандартного. Такое переосмысление приводит к компактизации геномов органелл. Подробности в предыдущем п., в частности см. табл. 4. В дополнение к утверждению в п.1 по поводу девиантных кодов констатируем, что ни о какой роли «случайности» в переосмыслении кодонов говорить невозможно.

Заключение. Первая постановка задачи по математическому анализу феномена перекрывающихся генов относится к 1992г. Основные полученные результаты его были опубликованы в 1994-2004гг. в Докладах Академии наук (10 статей по разделам математика, молекулярная биология).

Постановка новых задач, которые возникли в ходе проведения этого исследования была направлена прежде всего на расширение подтверждающей базы двух фундаментальных выводов, сформулированных выше. Помимо решения новых задач, поставленных нами для перекрывающихся генов, это привело также к необходимости изучения областей ДНК, в которых структурные гены не являются перекрывающимися, а подчиняются принципу сформулированному как предположение в 1941г.: один ген отвечает за один белок [55]. Проникновение в такие области ДНК сегодня - это выход на первые позиции современных многосторонних исследований по целым геномам (человека и др.). Нами были поставлены некоторые задачи (по биоматематике), которые решаются в настоящее время.

Автор благодарит академика Т.М.Энеева за предложение обратиться к задачам молекулярной биологии, постоянное внимание к данной работе и многократные обсуждения. Автор благодарит академика О.Б.Лупанова за активную поддержку данных исследований на самом важном - начальном этапе их представления (2000г.).

СПИСОК ЛИТЕРАТУРЫ

1. Козлов Н.Н., Сюняев Р.А., Энеев Т.М. 1974. Гравитационное взаимодействие галактик. *Вестник АН СССР*. **7**. 50-61.
2. Eneev T.M., Kozlov N.N., Sunyaev R.A. 1973. Tidal Interaction of Galaxies. *Astron. & Astrophys*. **22**. 41-60.
3. Eneev T.M., Kozlov N.N. 1981. The problems of simulation of Planetary systems accumulation processes. *Advanced Space Research COSPAR.* **1**. 201-251.
4. Энеев Т.М., Козлов Н.Н. 1981. Модель аккумуляционного процесса формирования планетных систем. I. Численные эксперименты. *Астрономический вестник.* **Т15. 2**. 80-94. II. Вращение планет и связь модели с теорией гравитационной неустойчивости. *Астрономический вестник.* **15**(3). 131-141.
5. Козлов Н.Н. 1989. Исследование множественных конфликтов при компьютерном проектировании БИС. *Препринт ИПМ АН СССР*. **131.** 27.
6. Энеев Т.М., Козлов Н.Н. 1982. О новом методе численного моделирования эволюции сложных дискретных систем. *Доклады Академии Наук СССР.* 263. **4.** 820-824.
7. Козлов Н.Н. 1984. Метод виртуальных контактов. *ЖВМ и МФ*. **24**(2). 218-239.
8. Kozlov N.N., Kugushev E.I. 1993. Computer simulation of tRNA secondary structure folding. *CABIOS*. **9.** 253-258.
9. Козлов Н.Н., Кугушев Е.И., Энеев Т.М. 1998. Структурообразующие характеристики транскрипционного процесса. *Математическое моделирование*. **10**(6). 3-19.
10. Козлов Н.Н., Кугушев Е.И., Энеев Т.М. 2000. Параллельные вычисления при решении некоторых задач астрофизики и молекулярной биологии. *Математическое моделирование*. **12**(7). 65-70.
11. Козлов Н.Н., Кугушев Е.И., Энеев Т.М. 2003. *Компьютерное моделирование и анализ биологических систем.* Сб. 50 лет ИПМ: Направления, исследования и достижения. 40-41.
12. Козлов Н.Н.. 1994. Об особом способе записи генетической информации. *ДАН*. **337**(1). 158-161.
13. Yčas, M. 1969. *The biological code*. Amsterdam. L. 359.
14. Ичас М. 1994. *О природе живого: механизмы и смысл*. М.: Мир. 496.

15. Mendel G. 1866. Versuche über Pflanzenhybriden, Verhandl. Naturforsch. Ver. Brünn. **4.** 3-47.
16. Шредингер Э. 1972. *Что такое жизнь? С точки зрения физика*. М.: Атомиздат. 88.
17. Watson J.D., Crick F.H.C. 1953. A structure for Deoxyribose Nucleic Acid. *Nature*. **171.** 737-738.
18. Уотсон Д. 1969. *Двойная спираль. Воспоминания об открытии структуры ДНК.* М.: Мир. 152.
19. Льюин Б. 1987. *Гены.* М.: Мир. 544.
20. Barrell B.G., Air G.M. and Hutchison C.A. 1976. III. Overlapping genes in bacteriophage ФХ174. *Nature.* **264.** 34-41.
21. Sanger F., Coulson A.R., Friedmann T., Air G.M., Barrell B.G., Brown N.L., Fiddes J.C., Hutchison C.A., III, Slocombe P.M., Smith M. 1978. The Nucleotide Sequence of Bacteriophage ФХ174. *J. Mol. Biol.* **125.** 225-246.
22. Ноздрачев А.Д., Поляков Е.Л., Зеленин К.Н. 2004. Первая нобелевская премия России. *Вестник РАН.* **8**.
23. Seeger C., Ganem D., Varmus H.E. 1984. Nucleotide Sequence of an Infectious Molecularly Cloned Genome of Ground Squirrel Hepatitis Virus. *J. Virol.* **51.** 367-375.
24. Testut P., Renard C-A., Terradillos O., VitvitskiTrepo L., Tekaia F., Degott C., Blake J., Boyer B., Buendia M.A. 1996. A New Hepadnavirus Endemic in Arctic Ground Squirrels in Alaska. *J. Virol.* **70.** 4210-4219.
25. Козлов Н.Н.. 1999. О востребованности каждого из 64 кодонов в генетических перекрытиях. *ДАН.* **367**(4). 544-547.
26. Козлов Н.Н.. 1996. Молчащие мутации в области перекрывания генов. *ДАН.* **350**(5). 699-703.
27. Козлов Н.Н. 1998. Терминаторные кодоны в генетических перекрытиях. *ДАН.* **360**(4). 550-553.
28. Schwartz D., Tizard R., Gilbert W. 1983. Nucleotide Sequence of Rous Sarcoma Virus. *Cell* (Cambridge, Mass). **32.** 853-869.
29. Kozlov N.N. 1996. A Theorem for overlapping genes. *Preprint Keldysh Institute of Appled Mathematics*. **115**. 23.
30. Godson G.N., Barrell B.G., Staden R., Fiddes J.C. 1978. Nucleotide Sequence of Bacteriophage G4 DNA. *Nature* (London). **276**. 236-247.
31. Wang L.F., Michalski W.P., Yu M., Pritchard L.I., Crameri G., Shiell B., Eaton B.T. 1998. A Novel P/V/C Gene in a New Member of the Paramyxoviridae Family, Which Causes Lethal Infection in Humans, Horses, and Other Animals. *J. Virol.* **72**(2). 1482-1490.
32. Козлов Н.Н. 1997. Перекрывающиеся гены и генетический код. *ДАН*. **355**(6). 830-833.
33. Козлов Н.Н. 1999. К вопросу о произвольности «выбора» генетического кода. *ДАН*. **369**(4). 553-556.
34. Anderson S., de Bruijn M., Coulson A. R., Eperos I. C., Sanger F., Young G. 1982. Complete Sequence of Bovine Mitochondrial DNA. *J. Mol. Biol.* **156.** 683-717.
35. Kröger M., Hobom G. 1982. Structural analysis of insertion sequence IS5. *Nature*. **297.** 159-162.
36. Rak B., von Reutern M. 1984. Insertion element IS5 contains a third gene. *The EMBO Journal*. **3**(4). 807-811.
37. Козлов Н.Н. 2004. Элементарные генетические перекрытия. *Препринт ИПМ им. М.В.Келдыша*. *РАН*. **64**. 27. http://www.keldysh.ru/papers/2004/prep64/prep2004_64.html.
38. Козлов Н.Н. 2002 Теорема для генетического кода. *ДАН*. **382**(5), 593-597.

39. Козлов Н.Н. 2000. Анализ полного множества перекрывающихся генов. *ДАН*. **373(1)**. 108-111.
40. Козлов Н.Н. 2004. Применение теоремы для генетического кода. *ДАН*. **396**(6). 740-745.
41. Jukes T.H. 1990. Genetic code 1990. *Outlook. Experientia.* **46.** 11-12.
42. Wang Lei, Brock Ansgar, Herberich Brad, Schultz Peter G. 2001. Expanding the genetic code of *Escherichia coli*. *Science*. **292**(5516). 498-500.
43. Barrell B.G., Bankier A.T., Drouin J. 1979. A different genetic code in human mitochondria. *Nature*. **282.** 189-194.
44. Anderson S., Bankier A.T., Barrell B.G., de Bruijn M.H.L., Coulson A.R., Drouin J., Eperon I.C., Nierlich D.P., Roe B.A., Sanger F., Schreier P.H., Smith A.J.H., Staden R. and Young I.G. 1981. Sequence and organization of the human mitochondrial genome. *Nature*. **290.** 457-464.
45. Crozier R.H., Crozier Y.C. 1993. The Mitochondrial Genome of the Honeybee *Apis mellifera*: Complete Sequence and Genome Organization. *Genetics* .
46. Cantatore P., Roberti M., Rainaldi G., Gadaleta M.N. Saccone C. 1989. The Complete Nucleotide Sequence, Gene Organization, and Genetic Code of the Mitochondrial Genome of *Paracentrotus lividus*. *The J. Biological Chemistry*. **264**(19). 10965-10975.
47. Gadaleta G., Pepe G., De Candia G., Quagliariello C., Sbisa E., Saccone C. 1989. The Complete Nucleotide Sequence of the *Rattus norvegicus* Mitochondrial Genome: Cryptic Signals Revealed by Comparative Analysis between Vertebrates. *J. Mol. Evol*. **28.** 497-516.
48. Козлов Н.Н. 2000. Перекрывающиеся гены и вариабельность генетического кода. *ДАН*. **375**(6). 824-827.
49. Clary D.O., Wolstenholme D.R. 1985. The Mitochondrial DNA Molecule of *Drosophila yakuba*: Nucleotide Sequence, Gene Organization, and Genetic Code. *J. Mol. Evol*. **22.** 252-271.
50. Smith M.J., Banfield D.K., Doteval K., Gorski S., Kowbel D.J. 1990. Nucleotide Sequence of Nine Protein-Coding Genes and 22 tRNAs in the Mitochondrial DNA of the Sea Star *Pisaster ochraceus*. *J. Mol. Evol*. **31.** 195-204.
51. Rees A.R., Sternberg M.J.E. 1984. *From Cells to Atoms. An Illustrated Introduction to Molecular Biology.*
52. Alberts B., Bray D., Lewis J., Raff M., Roberts K., Watson J. 1994. *Molecular Biology of the Cell.* New Jork, London: Gorland Publishing, Inc. 1294.
53. Энеев Т.М. 1970. *Некоторые вопросы применения метода наискорейшего спуска*. М.: Препринт ИПМ АН СССР. **17**.
54. Козлов Н.Н. 1995. Математический анализ особого способа записи генетической информации. *Математическое моделирование*. **7**(12). 33-47.
55. Beadle G.W., Tatum E.L. 1941. Genetic control of biochemical reactions in *Neurospora*. *Proc. Natl. Acad. Sci. USA*. **27**. 499-506.

Материал поступил в редакцию 28 февраля 2006 г., опубликован 29 марта 2006 г.

ПРИЛОЖЕНИЕ 1

Таблица 1. Структура для 4-х природных генетических кодов: для стандартного K^0 и нестандартных K^1-K^3, для которых приведены лишь переосмысленные кодонные семейства (см. ниже).

		K^0		K^1		K^2		K^3	
1	Met	(1)	ATG	(2)	ATX	(2)	ATX		
2	Trp	(1)	TGG	(2)	TGX	(2)	TGX	(2)	TGX
3	Phe	(2)	TTY						
4	Tyr	(2)	TAY						
5	His	(2)	CAY						
6	Asn	(2)	AAY					(3)	AAM
7	Asp	(2)	GAY						
8	Cys	(2)	TGY						
9	Gln	(2)	CAX						
10	Lys	(2)	AAX					(1)	AAG
11	Glu	(2)	GAX						
12	Ile	(3)	ATM	(2)	ATY	(2)	ATY		
13	Val	(4)	GTN						
14	Pro	(4)	CCN						
15	Thr	(4)	ACN						
16	Ala	(4)	GCN						
17	Gly	(4)	GGN						
18	Ser	(6)	TCN, AGY			(8)	TCN, AGN	(8)	TCN, AGN
19	Leu	(6)	CTN, TTX						
20	Arg	(6)	CGN, AGX	(4)	CGN	(4)	CGN	(4)	CGN
	ter	(3)	TAX, TGA	(4)	TAX, AGX	(2)	TAX	(2)	TAX

Примечание. При записи 20-и аминокислот были использованы стандартные трехбуквенные сокращения. Для каждой из аминокислот приводятся общепринятые трехбуквенные сокращения. Для стандартного кода K^0 указано число кодонов-синонимов (в скобках) и их трехбуквенные представления. Обозначения: X: A, G; Y: T, C; M: T, C, A; N: A, G ,T, C. В последней строке приводятся три терминаторных кодона - ter, каждый из которых обозначает останов синтеза белка.

Таблица 2. Частота встречаемости кодонов в областях, кодирующих перекрывающиеся гены в бактериофаге ФХ174.

Phe	TTT	11	Ser	TCT	5	Tyr	TAT	7	Cys	TGT	5
	TTC	13		TCC	5		TAC	8		TGC	4
Leu	TTA	15		TCA	7	Ter	TAA	0	Ter	TGA	4
	TTG	11		TCG	9		TAG	0	Trp	TGG	6
Leu	CTT	12	Pro	CCT	8	His	CAT	3	Arg	CGT	7
	CTC	8		CCC	2		CAC	2		CGC	11
	CTA	5		CCA	3	Gln	CAA	13		CGA	6
	CTG	19		CCG	6		CAG	10		CGG	5
lle	ATT	16	Thr	ACT	13	Asn	AAT	15	Ser	AGT	5
	ATC	8		ACC	6		AAC	14		AGC	5
	ATA	3		ACA	7	Lys	AAA	27	Arg	AGA	7
Met	ATG	15		ACG	11		AAG	14		AGG	1
Val	GTT	10	Ala	GCT	17	Asp	GAT	10	Gly	GGT	2
	GTC	9		GCC	5		GAC	13		GGC	6
	GTA	4		GCA	8	Glu	GAA	18		GGA	9
	GTG	7		GCG	12		GAG	12		GGG	3

Таблица 3. Структура шести гипотетических кодов образованных из K^0. Первая пара - K_{11}, K_{12} - образована при помощи перестановки всего одного смыслового кодона. Паре K_{21}, K_{22} соответствует характеристика *p*=0. Расчет K_{31}, K_{32} осуществлялся с целью увеличения значения *p* вплоть до почти десятикратного по сравнению с p^0 (это *p* для K^0).

	K^0 (4)	K_{11} (4)	K_{12} (4)	K_{21} (4)	K_{22} (0)	K_{31} (3)	K_{32} (4)
Met	(1) ATG				(4) ATN		
Trp	(1) TGG	(2) TGG,TGC			(2) TGX	(1) TAG	(1) TTG
Phe	(2) TTY				(4) TTN		(1) CTG
Tyr	(2) TAY			(3) TAY,TGA	(3) TAM		(1) GTG
His	(2) CAY						(1) CAT
Asn	(2) AAY						
Asp	(2) GAY						
Cys	(2) TGY	(1) TGT				(1) TAA	(1) CAC
Gln	(2) CAX						(1) CAG
Lys	(2) AAX						
Glu	(2) GAX						(1) CAA
Ile	(3) ATM				(4) AGN		
Val	(4) GTN						(5) GTM,GAX
Pro	(4) CCN						(8) CCN,TAN
Thr	(4) ACN						
Ala	(4) GCN						
Gly	(4) GGN						
Ser	(6) TCN,AGY				(4) TCN		(4) AGN
Leu	(6) CTN,TTX				(4) CTN		(6) YTM
Arg	(6) CGN,AGX		(5) CGN_A,AGX		(4) CGN		(8) CGN,TCN
ter	(3) TAX,TGA		(4) TAX,YGA	(2) TAX	(1) TAG	(4) TGN	(4) TGN
p	16	22	28	0	0	58	155

Примечание: помимо характеристики *p* (нижняя строка) в скобках верхней строки дано число нерегулярных кодонных семейств; оно нулевое только для K_{22} - у этого кода всего один кодон терминации. Обозначения: N: A, C, T, G; N_A: C, T, G; M: A, C, T; X: A, G; Y: T, C.

Таблица 4. Эффект сокращения размера митохондриальных ДНК (для 6 геномов, размер каждой ДНК ~16500 пар нуклеотидов) за счет переосмысления кодонов в нестандартных кодах K^{1-} K^3 (см. табл.1) Количество митохондрий в одной - единственной клетке может достигать нескольких тысяч [51]

1	2	3	4
Human [44]	K^1	TGA (Trp)	45(8)
R.norvegicus [47]	K^1	TGA (Trp), ATA (Met)	73(8)
D. Yakuba [49]	K^2	TGA (Trp), ATA (Met)	7(0)
A. Mellifera ligustica [45]	K^2	TGA (Trp), ATA (Met)	19(0)
P. Lividus [46]	K^3	TGA (Trp), AAA (Asn)	16(0)
P. ochraceus [50]	K^3	TGA (Trp)	9(0)

Примечание: столбец 1- митохондриальный геном, 2- нестандартный код, 3 - переосмысленные кодоны, которые делают возможными перекрытия генов для нестандартного кода, или кодоны, переводящие БРС для K^0 в ОРС для нестандартного кода, 4 - размер сокращения ДНК (пары нуклеотидов) за счет переосмысления кодонов; в скобках указывается размер перекрытий, которые не зависят от переставленных кодонов и существуют как для нестандартного кода, так и для стандартного. Для случая *P. ochraceus* данные столбцов 3,4 получены только по последовательности 8028 нуклеотидов согласно рис. 2 из [50].

ПРИЛОЖЕНИЕ 2

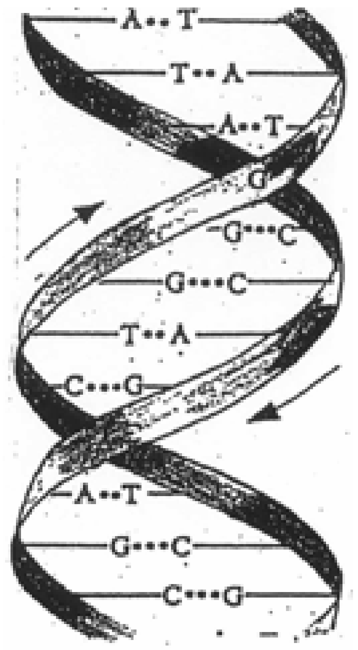

Рис. 1. Модель двойной спирали ДНК. Рисунок создан на основе самого первого рис. из [17]. Алфавит ДНК содержит всего 4 буквы A (это нуклеотид аденин), C (- цитозин), G (- гуанин), T (-тимин). Между спиралями существуют только связи A с T (число водородных связей - две) и C с G (- три). Чтение текста гена указано стрелками по одной цепи - сверху вниз, по другой - снизу вверх.

```
сдвиг -1  AsnGlyGlyLeuLeu *  * AlaGlu...          БРС
сдвиг +1    TrpArgLeuArgIleValGly * ...           БРС

белок →    MetGlnAlaCysTyrSerArgLeuLys...         ОРС
ген →     AATGGAGGCTTGCTATAGTAGGCTGAAG...
           •               ↑   ↑    ↑
                           C   C    C
```

Рис. 2. Участок белковой последовательности (первая аминокислота - Met) закодирован в гене начиная с ATG (первый нуклеотид A в этом триплете помечен жирной точкой). Этот участок соответствует открытой РС - ОРС. Генетики показали, что для типичного гена при сдвиге начальной точки (чтения гена) на +1 либо -1 нуклеотид получим другие последовательности кодонов (другие РС), в каждой из которых будет присутствовать кодон-терминатор ter - на рис. помечен символом *. Это будут две РС с блокировками - БРС. С помощью трех нуклеотидных замен на нуклеотид C (указано под текстом гена) ни один из трех кодонов ter (символ *) на заданном участке гена не возникнет, причем при таких заменах белковая последовательность не изменится т.к. указанные три замены соответствуют трем заменам кодонов на их синонимы. Однако типичный ген устроен так, чтобы указанные сдвиги давали именно две БРС [19].

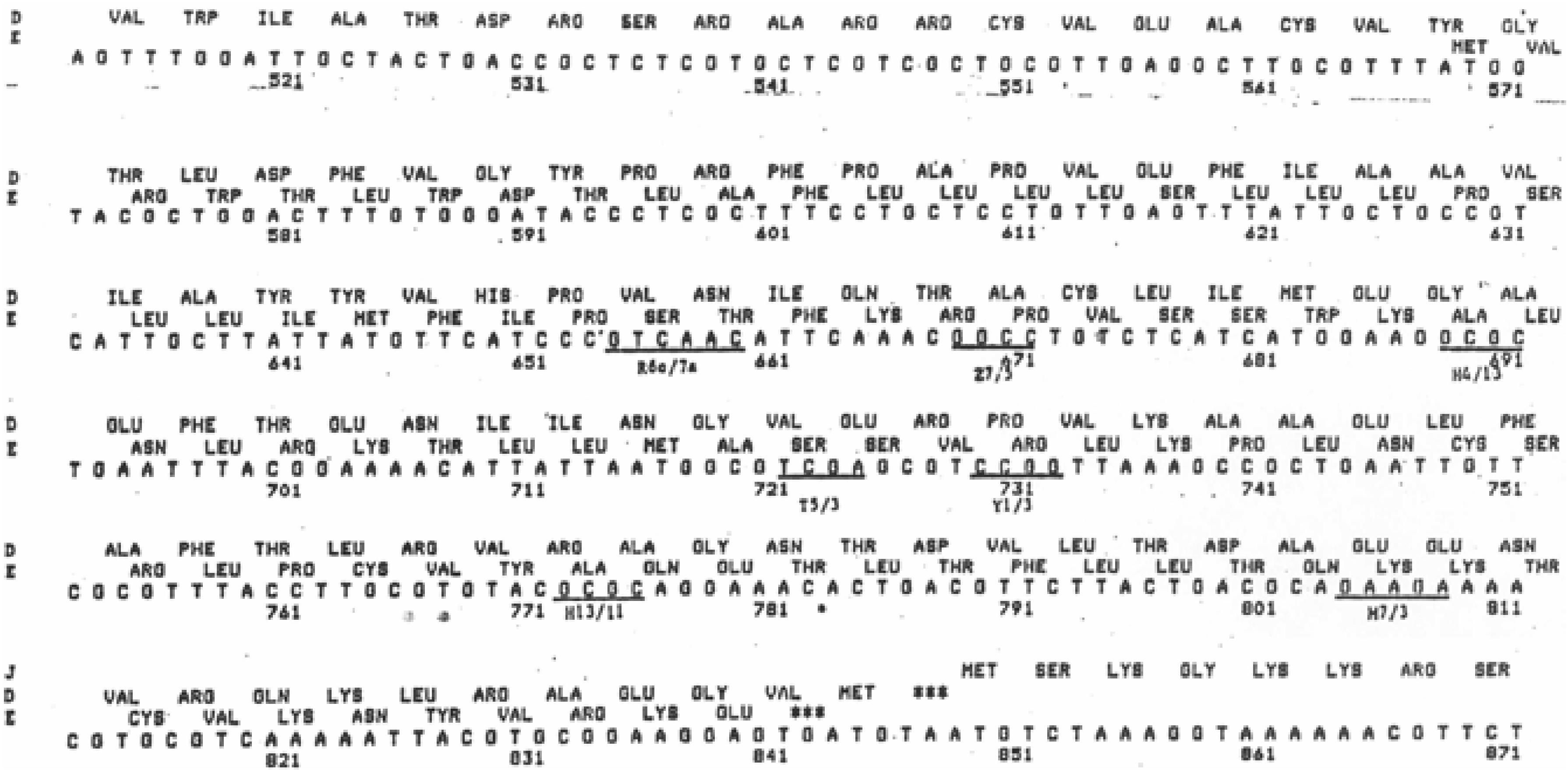

Рис. 3. Первое из перекрытий генов, обнаруженных в 1976г.: РС белка E сдвинута на +1 нуклеотид относительно ОРС для белка D. Ген белка E начинается с позиции 567 и целиком расположен внутри гена D. Рисунок соответствует полному тексту генома ФХ174 из [21]; первая публикация [20]. Оказалось, что суммарный размер перекрытий в этом геноме составляет около 16% размера генома - 814 нуклеотидов, из которых 4-е нуклеотида отвечают перекрытию трех генов [21].

```
A   GlyTyr     AlaAsp     A     Leu     K     Leu
B     Leu         Arg     K   ValTyr    C   SerTyr
    GGTTA      GCCGA        GTTTA         TCTTA
      C           A            C             C
   5216       5300           98           174


D     Leu      Leu             Leu           Leu
E   ArgTrp   AsnCys          ProCys        AsnTyr
    CGCTG    AATTG           CCTTG         AATTA
      T        C               C             C
    576      747             762           825


D       ValMet
E        ter
        GTGA
          A
        842
```

Рис. 4. Полный перечень локальных перекрытий из ФХ174, в которых допустимы нуклеотидные замены, соответствующие молчащим мутациям [26]. Названия белков (слева): A, B, C, D, E, K и нумерация нуклеотидов соответствуют [21]. Над порядковыми номерами нуклеотидов в геномах указаны нуклеотиды - замены, соответствующие молчащим мутациям. В данном случае могут быть использованы 17 смысловых кодонов-синонимов (только они могут повлиять на численность используемых кодонов в табл. 2 при любых вариантах записи перекрытий): 7 кодонов Leu, по 2 кодона для трех аминокислот: Arg, Asn, Val и по одному кодону для четырех аминокислот: Ser, Gly, Ala, Pro. С учетом этих данных число используемых кодонов (61) может уменьшиться лишь за счет кодонов Leu - CTA (их 5 в таблице 2) и кодона Arg - AGG (такой кодон всего 1), т.к. частота встречаемости каждого из кодонов Asn, Val больше двух, а кодонов Ser, Gly, Ala, Pro - больше 1 (см. табл. 2). Расчеты показали, что при любых допустимых кодонах-синонимах полное число кодонов в записи перекрытий в ФХ174 не может быть менее 61 или полного числа смысловых кодонов в K^0.

```
       G4                            HeV
  D     ter          SB   ter              V      ter
  E   CysLys         P   ValVal            P    AsnGlu
        TAA                TAG                    TGA
     2431                938                  1476
```

Рис. 5. Три локальных перекрывания из G4 [30] и HeV [31]. Они содержат терминаторные кодоны TAA, TAG, TGA, которые не могут быть заменены кодонами-синонимами. Названия белков даны слева.

Перекрытия генов из одной цепи ДНК

сдвиг -1

B_{11}→... LysIleTyr...
B_{12}→...GluAsnLeuPhe...
...GAAAATTTATTT...

1

сдвиг +1

B_{21}→ MetLeu...
B_{22}→...GlnCys...
...CAATGCTA...

2

Перекрытия генов из разных цепей ДНК

сдвиг -1

B_{31}→...ValAsnAsn...
...GTGAATAACA...
...GCACTTATTGTA...
...ThrPheLeuMet...←B_{32}

3

сдвиг 0

B_{41}→ MetLeuVal
...ATGCTCGTT...
...TACGAGCAA...
...HisGluAsn...←B_{42}

4

сдвиг +1

B_{51}→...AsnPheHisGlu...
...AATTTCCACGAG...
...TTAAAGGTGCTC...
...AsnGlyArg... ←B_{52}

5

Рис. 6. 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}. Сдвиг между генами равен -1 нуклеотид для случаев 1, 3 и +1 - случаев 2, 5; в случае 4 подобный сдвиг равен 0. В отличие от рис. 1 из [38] в данном рис. все случаи перекрытий соответствуют только двум геномам; автору не известен геном, где используются все 5 случаев перекрываний. Случаи 1, 2, 5 присутствуют в кодировке mtДНК Bovin [34], а случаи 3, 4 - IS5 [35, 36]. Отметим, что белки в геноме mtДНК Bovin закодированы девиантным кодом (см. также ниже), однако во всех трех приведенных фрагментах перекрытий участвуют только кодировки, соответствующие K^0. Укажем начальные позиции для перекрываемых фрагментов, а также названия белков. Для mtДНК Bovin [34] имеем: B_{11} соответствует URFA6L (первый нуклеотид в указанном перекрытии - 8297), B_{12} - ATPaзe6; B_{21} - URF4(-10529), B_{22} - URF4L, B_{51} - URF5 (- 13915), B_{52} - URF6. Для IS5 [35, 36] имеем: B_{31} - ins 5C(-205), B_{32} и B_{42} являются фрагментами ins 5A, а B_{41} - ins 5B (-525).

```
              1      2      3         80

            Met    Met    Met   ...  Arg
  W1(80)   Tyr    His    Asn         Ser
           TATG   CATG   AATG        TCGN


              1      2      3          80

            Met    Met    Trp   ...  Arg
  W2(80)     Trp    Cys    Gly         Gly
            ATGG   ATGY   TGGN        ZGGN


              1      2      3          35

            Met    Met    Trp          Arg
  W3(35)    ATG    ATG    TGG    ...   AGX
           GTA    MTA    YAC         NTC
           Met    Ile    His         Leu


              1      2      3          52

            Met    Trp    Phe          Arg
  W4(52)    ATG    TGG    TTT    ...   CGC
            TAC    ACC    AAA          GCG
            His    Pro    Lys          Ala


              1      2      3          201

            Met    Met    Met          Arg
  W5(201)   ATG    ATG    ATG    ...   AGG
             ACC    ACA    ACG          CCT
             Pro    Thr    Ala          Ser
```

Рис. 7. Некоторые э.п. из W_1-W_5. 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 `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`.

		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

↑ AA^l AA^t ⟶

Рис. 8. Участие аминокислот в элементарных перекрытиях - э.п. – из множеств W_1-W_5. По оси абсцисс отложена аминокислота из верхних строк э.п. - AA^t; t - top (верхний), а по оси ординат - аминокислота из нижних строк э.п. (это вторая строка в э.п. из W_1, W_2 и четвертая - для э.п. из W_3-W_5). - AA^l, l - lower (нижний). Из 400 возможных позиций только заштрихованные 295 позиций заняты одним или более э.п. Помимо позиций, занятых числами 1-5, которые соответствуют одиночным э.п. из W_1-W_5 двухзначными числами представлены также 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. указажем далее 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).

```
                                                                 GSHV
  3      57      19        12      37
 Met    Gly     Asn       Tyr     Ile →        MetGlyAsn...TyrIleter  S
Asn    Gly    Gln       Ile     His      ...  AsnGlyGln...IleHis...   A
AATG  GGGN   CAAY ...  ATAY   CATM      ... AATGGGCAA...ATACATCTAA
  1      2      3        427     428         .                .
                                          1059             2341
```

Рис. 9. Пример сборки (или стыковки) элементарных перекрытий - э.п. (номера их 3, 57, 19, 12, 37 соответствуют множеству W_1 из [33] для перекрытия генов S и A в GSHV содержащего кодонов (номера их приведены под э.п.). Для того чтобы перекрытия на основе э.п. стало возможным должно выполняться условие: первый нуклеотид (или их набор) последующего э.п. должен содержаться в четвертой позиции (где может быть как один нуклеотид, например э.п. 57 содержит N в четвертой позиции; N включает в себя C соответствующее первой позиции э.п. 19). Отметим, что подобное условие не выполняется даже для всех возможных перекрытий для пар аминокислот: множество перекрытий для пар аминокислот допускает лишь 4695 перекрытий из возможных 6400. На рисунке подчеркиванием отмечены объединяемые позиции для конкретного перекрытия генов S и A. Для построения полного множества перекрытий генов в процессе указанной сборки следует учитывать все допустимые стыковки. Этот учет показал [33], что число возможных различающихся перекрытий для 428 кодонов ~ 10^{746}.

```
                        1.ter: TGA

         →           →           →           →           →
PC O     MetMet      MetAsn      MetLys      MetIle      MetThr
         ATGATG      ATGAAY      ATGAAX      ATGATM      ATGACN
PC 2      ter         ter         ter         ter         ter
          →           →           →           →           →


                      2.ter: TAA, TAG

         →           →           →           →           →           →
PC O     PheTyr      TyrTyr      HisTyr      AsnTyr      AspTyr      CysTyr
         TTYTAY      TAYTAY      CAYTAY      AAYTAY      GAYTAY      TGYTAY
         AAXAT       ATXAT       GTXAT       TTXAT       CTXAT       ACXAT
PC 3       ter         ter         ter         ter         ter         ter
           ←           ←           ←           ←           ←           ←


                      3.ter: TAA, TGA

         →           →           →           →           →
PC O     PheMet      PheAsn      PheLys      PheIle      PheThr
         TTYATG      TTYAAY      TTYAAX      TTYATM      TTYACN
         AAXTAC      AAXTT       AAXTT       AAXTA       AAXTG
PC 5      ter         ter         ter         ter         ter
          ←           ←           ←           ←           ←
```

Рис. 10. Перекрытия для 16 пар аминокислот из PC0, для которых согласно теореме из [38] имеет место неизбежное возникновение БРС (кодоны терминации выделены жирным шрифтом). Для фрагментов 2. и 3. гены берутся из различных цепей ДНК, для фрагмента 1. - из одной.

1. PC2 сдвинута на +1 нуклеотид относительно PC0. Неизбежное возникновение ter: TGA имеет место для каждой из 5-и пар из (1).
2. PC3 сдвинута на -1 нуклеотид относительно PC0 блокировка PC3 возникает из-за ter: TAA или TAG для каждой из 6-и пар аминокислот из (2).
3. PC5 сдвинута на +1 нуклеотид относительно PC0. Из-за ter: TAA либо TGA имеет место блокировка PC5 для каждой из 5-и пар а.к. из (3).

Полное число пар *p* для K^0 оказалось равным 16: по 5 для PC2 и PC5 и 6 для PC3. Показывается [38], что для K^0 никаких других блокировочных пар не существует. Приведены направления чтения для PC. Оказалось также, что все 3 кодона терминации участвуют в какой-либо БРС: кодон TGA (в PC 2 и PC 5), кодон TAA (в PC3 и PC5), а кодон TAG участвует в блокировке всего одной PC (PC3).

K^1 (Human)

ATPase6→ **MetAsn**GluAsnLeuPheAlaSerPheIleAlaProThrIleLeuGly...
URF A6L→ ...Lys**Trp**ThrLysIleCysSerLeuHisSerLeuProProGlnSerter
...AAA**TGA**ACGAAAATCTGTTCGCTTCATTCATTGCCCCCACAATCCTAGGCC...
. . . . .
8530 8570

K^1 (Rattus norvegicus)

ATPase6→ **MetAsn**GluAsnLeuPheAlaSerPheIleThrProThr**Met**Met...
ATPase8→ ...Lys**Trp**ThrLysIleTyrLeuProLeuSerLeuProProGlnter
...AAA**TGA**ACGAAAATCTATTTGCCTCTTTCATTACCCCCACA**ATA**ATA...
. . .
7910

terAsnArgThrIleGluIleIleIlePhe...←ND6
ND5→...LeuAsnProGlu**Trp**PheGlnterterter
...CTTAATCCCGAG**TGA**TTTCAATAATAATAAA...
. . .
13530

K^2 (Apis mellifera ligustica)

ATPase6→ **MetLys**LeuIleLeu**Met**Met...
ATPase8→ ...Lys**Trp**Asn**Trp**Phe**Trp**ter
... AAA**TGA**AAT**TGA**TTT**TGATA**ATA...
. .
4590

K^3 (Paracentrotus lividus)

ATPase6→ **MetThrMetThr**IleThr...
ATPase8→ ...**AsnTrp**Gln**Trp**Leuter
... **AAATGA**CAA**TGA**CTATAACTG...
.
8680

Рис. 11. Перекрытия из митохондриальных ДНК четырех организмов, записанных нестандартными K^1-K^3 кодами соответственно. Верхний рисунок соответствует перекрытию в митохондриальной ДНК человека (впервые [48]). Приведены лишь фрагменты, где использованы переосмысленные кодоны, а размеры перекрытий без участия таких кодонов указаны в скобках в стобце 4 табл. 4. Жирным шрифтом отмечены переосмысленные кодоны TGA(Trp), ATA(Met), AAA(Asn), а также блокировочные пары из (1): MetAsn(K^1), MetLys(K^2) и дважды MetThr(K^3). Для подобной пары MetMet(K^2) было обнаружено перекрытие у D.Jakuba [49] (фрагмент его представлен на рис. 2 из [38]). Для K^1 (R.norvegicus [47]) приведены два указанных перекрытия генов, принадлежащих одной цепи ДНК (сдвиг -1, первый фрагмент) либо двум цепям ДНК (сдвиг +1, второй фрагмент, РСО соответствует гену ND5). Для K^2 (A.Mellifera ligustica [45]) и K^3 (P.lividus [46]) приведены только по одному фрагменту.