

© 1990 г.

ГЕЛЬФАНД М. С.

КОДЫ ГЕНЕТИЧЕСКОГО ЯЗЫКА И ЕСТЕСТВЕННЫЙ ЯЗЫК

Тремя основными открытиями, каждое из которых послужило началом очередного этапа в развитии молекулярной генетики, были определение Уотсоном и Криком структуры ДНК (1953 г.), расшифровка генетического кода (начало 60-х годов) и разработка быстрых методов секвенирования (определения) нуклеотидных последовательностей (1977 г.). Первое из них позволило установить, как хранится и воспроизводится генетическая информация, второе — каким образом осуществляется ее экспрессия (декодирование), третье привело к лавинообразному росту числа известных последовательностей и, в частности, позволило приступить к детальному изучению регуляции молекулярно-генетических процессов.

Примерно в это же время в работах лингвиста Р. Якобсона и биолога Ф. Жакоба [1, 2] были впервые обсуждены наблюдаемые аналогии между языковыми и генетическими структурами. В частности, лингвистами была выдвинута гипотеза о филогенетической связи языка и генетического кода. По словам Т. В. Гамкрелидзе, «якобсоновское понимание структурного изоморфизма между генетическим и лингвистическим кодами предполагает эволюционный процесс наложения лингвистического кода непосредственно на генетический и копирования его структурных принципов, осуществившийся в условиях бессознательного владения живым организмом знанием характера и структуры последнего... Это выразилось не только в филогенетическом процессе оформления структур языкового механизма по модели генетического кода, но и в различных творческих актах отдельных выдающихся личностей, строящих особые информационные (семиотические) системы в общем по модели генетического кода без эксплицитного знания структуры последнего» [3]. При этом основным (и практически единственным) доводом в пользу существования этой зависимости является особая значимость в этих семиотических системах соединения четырех различных элементов в тройки, что изоморфно таблице генетического кода (см. ниже).

С другой стороны, биологи часто используют лингвистические аналогии при описании молекулярно-генетических явлений, в частности, популярны сравнения тех или иных регуляторных участков в молекулах белков и нуклеиновых кислот с различными частями речи [4] или с различными видами предложений [5].

Оба эти подхода кажутся нам не вполне правильными. Прямое применение лингвистических аналогий в молекулярной генетике часто бывает поверхностным в том смысле, что используемые при этом лингвистические факты не универсальны — и поэтому аналогия не может быть глубокой. С другой стороны, решающими аргументами при рассмотрении вопроса о филогенетической зависимости естественного языка от языка генетического могли бы быть такие черты сходства между этими семиоти-

ческими системами, которые выделяли бы их среди прочих, — однако наши представления о генетическом языке настолько неполны, что ставить задачу нахождения этих специфических особенностей было бы, на наш взгляд, несколько преждевременно.

Поэтому в настоящей работе мы попытаемся применить несколько иной подход. Мы будем исходить из того, что и естественный язык, и язык генетический являются примерами больших и сложно организованных информационных систем совершенно различной природы (это нулевая гипотеза, не требующая априорных предположений). Поэтому их сопоставление может способствовать пониманию общих принципов построения таких систем, коль скоро эти принципы существуют. В связи со сказанным выше мы будем крайне осторожно подходить к поиску прямых аналогий. В то же время мы попытаемся описать набор фрагментов генетического языка, который иллюстрировал бы основные явления этого языка и был достаточен для построения на его основе каких-либо умозаключений.

Генетическим языком мы будем называть правила взаимодействия белков и нуклеиновых кислот между собой и друг с другом, и в частности правила, определяющие пространственную структуру этих макромолекул. Сами макромолекулы являются текстами этого языка. Описание некоторых основных явлений генетического языка содержится в п. 1—2.

Здесь следует сделать оговорку. Подобно тому, как слово «язык» в зависимости от контекста может означать какой-либо конкретный язык или же совокупность языков в целом, так и «генетический язык» — это язык какого-либо вида (скорее, группы видов, не различающихся на молекулярно-биологическом уровне) либо множество всех генетических языков.

1. Биологическое введение. В этом пункте мы по необходимости кратко и грубо опишем строение белков и нуклеиновых кислот и некоторые основные молекулярно-генетические процессы. Более подробное популярное введение можно найти в [6], а профессиональное — например, в хорошо написанном учебнике Дж. Уотсона [7].

Молекула ДНК (дезоксирибонуклеиновой кислоты) — полимер, с регулярным остовом которого связаны нуклеотиды, обозначаемые А, С, G и Т. В чередовании этих четырех нуклеотидов закодирована вся генетическая информация. Пространственная структура ДНК — двучечная спираль, цепи которой связаны друг с другом химическими связями между так называемыми комплементарными нуклеотидами $A = T$ (две водородных связи) и $G \equiv C$ (три водородных связи). Поэтому одна цепь может быть однозначно восстановлена по другой (в частности, благодаря этому осуществляется репликация — копирование молекулы ДНК). На цепях задано направление считывания, в котором проходят все биохимические процессы; при этом направления считывания комплементарных цепей противоположны. В клетке имеется одна основная молекула ДНК; ДНК разных клеток одного организма в общем тождественны.

У эукариотов (организмов, клетки которых имеют ядро) ДНК содержится в ядре. Организмы, не имеющие ядра, называются прокариотами; все они одноклеточны. ДНК человека (типичный многоклеточный эукариот) имеет длину примерно три миллиарда нуклеотидных пар (т. е. по три миллиарда нуклеотидов в каждой цепи), ДНК кишечной палочки (типичный прокариот) — три миллиона нуклеотидных пар, ДНК вирусов и фагов (прокариотных вирусов) — от нескольких тысяч до сотен тысяч нуклеотидных пар.

Вторым типом нуклеиновых кислот является рибонуклеиновая кислота (РНК). Молекулы РНК состоят из другого остова и тех же четырех нуклеотидов, с той лишь разницей, что вместо Т в состав РНК входит У, который также спаривается с А двумя водородными связями. Одноцепочечные молекулы РНК длиной до нескольких тысяч нуклеотидов являются копиями соответствующих участков ДНК. При этом некоторые РНК независимо (например, тРНК — транспортные РНК) или в составе белково-нуклеиновых комплексов (например, рРНК — рибосомальные РНК) осуществляют те или иные биохимические реакции, а информационные, или матричные, РНК (мРНК) кодируют белки. Процесс создания РНКовой копии участка ДНК называется транскрипцией и осуществляется белковым ферментом РНК-полимеразой; участок ДНК, который РНК-полимераза узнает перед началом работы, называется промотором, а сигнал к окончанию транскрипции — терминатором.

Отдельные участки РНК могут взаимодействовать с образованием двухцепочечных фрагментов. При этом так же, как в ДНК, спаренные фрагменты должны быть комплементарны и направлены противоположно друг другу, например, ССГТААГ и ГТТАСГГ; комплементарность не обязательно должна быть абсолютно строгой. Спаренные участки являются элементами вторичной структуры. Кроме чисто структурной роли, они могут быть важны для регуляции некоторых процессов.

Рибосомы, сложные белково-нуклеиновые комплексы, синтезируют белки, используя в качестве матриц молекулы мРНК. У прокариот на одной молекуле мРНК могут быть рядом закодированы несколько белков (участок ДНК, соответствующий такой матрице, называется оперон), у эукариот — как правило, один белок. Само кодирование осуществляется следующим образом. Каждому триплету нуклеотидов соответствует одна из 20 аминокислот, которые являются элементарными единицами, образующими белки. Таблица соответствий между тройками нуклеотидов (кодонами) и аминокислотами называется генетическим кодом. Считывая один кодон, рибосома присоединяет соответствующую аминокислоту (которая доставляется транспортной РНК) к растущему белку, а затем сдвигается к следующему кодону, не перекрывающемуся с прочитанным. Три кодона из 64 кодируют не аминокислоту, а сигнал окончания синтеза: встретив такой терминирующий кодон, рибосома высвобождает синтезированный белок и покидает матрицу. Этот этап белкового синтеза называется трансляцией.

2. Молекулярно-генетические коды. Из сказанного выше ясно, что генетический язык состоит из нескольких вообще говоря независимых кодов. Первое фундаментальное деление — на белковые и нуклеиновые коды. Мы сосредоточимся на кодах нуклеиновых кислот, хотя основные явления, которые мы собираемся описать, существуют и в белковых последовательностях.

Второе деление — на функциональные и структурные коды. Функциональные коды, как правило, связанные с белок-нуклеиновым взаимодействием, описывают регуляцию процессов репликации, транскрипции, трансляции и др., а структурные задают правила образования элементов структуры второго и последующих порядков; они в основном связаны с белок-белковым и нуклеино-нуклеиновым взаимодействием. Это разделение, как и сами термины, является до некоторой степени условным, как показывает пример, разобранный в п. 3.

Самым простым кодом является сам генетический код, упомянутый в предыдущем пункте. Он практически универсален (отклонения в отдельных организмах затрагивают небольшое число кодонов). Генетический код осуществляет соответствие между нуклеотидными и аминокислотными последовательностями и каждый триплетный кодон можно рассматривать как знак, означаемым которого является соответствующая аминокислота. При этом отдельный нуклеотид, входящий в состав кодона, сам по себе значением не обладает. Генетический код был также первой выясненной подсистемой генетического языка, и он первым обратил на себя внимание лингвистов, которые усмотрели уникальные структурные особенности, роднящие генетический код с естественным языком: «Среди всех систем передачи информации лишь генетический код разделяет со словесным кодом последовательное расположение дискретных субъединиц; сами по себе эти субъединицы — фонемы в языке и нуклеотиды (или „нуклеотидные буквы“) в генетическом коде — лишены врожденного значения, но они образуют те минимальные единицы, которые обладают своим собственным внутренним смыслом», и далее: «составление значащих единиц из дискретных подъединиц, лишенных своего собственного значения, объединяет только эти два кода из всех систем коммуникации» [1, с. 99, 101].

Однако уже здесь мы встречаемся с парадоксом, который, по-видимому, не имеет естественной лингвистической параллели. Дело в том, что аминокислоты, являющиеся означаемым для знаков генетического кода, — это элементарные единицы различных белковых кодов и тем самым они полностью аналогичны нуклеотидам и фонемам.

Структурные коды задают пространственное сворачивание нуклеиновых кислот сначала во вторичную (задаваемую спариваниями отдельных комплементарных участков), а затем в третичную структуру, а также взаимодействия между различными молекулами РНК или между РНК и белками, входящими в состав нуклеопротеиновых комплексов. Хотя пространственные структуры ряда молекул описаны при помощи рентгеноструктурных и биохимических экспериментов, никаких правил, описывающих структуры третьего порядка и выше, не известно. С другой стороны, вторичную структуру небольших молекул (например, тРНК) иногда удается предсказать физическими методами (грубо говоря, максимизируя число возникающих водородных связей), а для молекул большей длины, если известны их представители из многих организмов (например, рРНК), вторичная структура предсказывается сравнением многих возможных вторичных структур для каждой рРНК: правильна та вторичная структура, которая имеется у всех рассматриваемых рРНК. Сами последовательности при этом могут довольно существенно различаться, но тогда имеет место то, что называется комплементарными заменами: если два нуклеотида были спарены в каком-либо элементе вторичной структуры, то в рРНК другого организма они изменены оба так, что комплементарность между ними сохраняется. Как бы то ни было, имеется первоначальная классификация элементов вторичной структуры РНК и белков и постепенно накапливаются сведения о возможных их сочетаниях.

Наконец, третью группу составляют коды, связанные с узнаванием ДНК и РНК белками и рибонуклеопротеидами, осуществляющими или регулируемыми биохимические процессы или отдельные реакции. Здесь, пожалуй, правила наиболее разнообразны. Некоторые из этих правил совершенно строгие, в частности, правила узнавания сайтов (участков) разрезания ДНК рестриктазами (белками, осуществляющими этот процесс): например, рестриктаза EcoRI узнает

последовательность GAATTC, а рестриктаза EcoRII последовательности CCAGG и CCTGG и только их.

Другим примером сравнительно строгих правил, которые постепенно становятся, известны, может служить опознавание тРНК аминокислот-тРНК синтетазами (белками, присоединяющими к тРНК соответствующую ей аминокислоту). При этом в клетке имеется 20 aa-синтетаз (по одной на каждую аминокислоту), но многим аминокислотам соответствуют несколько различных тРНК, каждая из которых узнает свое подмножество кодонов. Однако все такие тРНК имеют некоторую структурную особенность (детерминанту), выделяющую их из остальных. Примечательно, что хотя первичные и особенно пространственные структуры различных тРНК близки, aa-синтетазы очень разнообразны и соответственно разнообразны детерминанты тРНК, которые могут быть расположены в совершенно различных частях молекулы (так что теоретически возможно сконструировать тРНК, которая узнавалась бы более чем одной aa-синтетазой).

Однако ситуация с большинством других белков и опознаваемых ими участков (сайтов) существенно менее ясна. Мы можем экспериментально определить примерное положение на ДНК распознающего белка и можем, меняя по одному нуклеотиды в этом участке, оценить существенность для распознавания отдельных позиций и стоящих в них нуклеотидов [6, рис. на с. 22] (Ю. С. Степанов обратил внимание автора на сходство этой экспериментальной процедуры с методами дескриптивной лингвистики). Мы можем также написать один под другим последовательности всех известных сайтов данного типа, выровняв их по какой-нибудь биологически выделенной точке (например, промоторы — участки старта транскрипции — выравнивают по точке начала транскрипции, а сайты старта трансляции — по иницирующему кодону, соответствующему первой аминокислоте). В результате этой процедуры можно обнаружить неравномерности в позиционных частотах нуклеотидов и сравнить их с результатами экспериментов по деактивации сайтов под воздействием мутаций. Обычно оказывается, что критические с экспериментальной точки зрения позиции имеют наиболее выраженное предпочтение одних нуклеотидов перед другими. Тем не менее, в подавляющем большинстве случаев создать правила, которые позволили бы строго выделять сайты среди всех последовательностей, не удается. Кажется весьма вероятным, что белки узнают какие-то структурные особенности, являющиеся тем самым элементарными единицами данного кода и не сводимые к признакам вида «нуклеотид в позиции p », которыми вынуждены пользоваться мы (и которые естественны для других кодов). Следует также иметь в виду, что разные сайты узнаются с разной эффективностью, которая по-разному зависит от биохимических условий, и поэтому скорее следует говорить не о разделении последовательностей на сайты (данного вида) и несайты, а об относительной эффективности разных участков последовательностей. Соответственно, смысл фрагмента последовательности не сводится к узнаванию его белком, а включает константу связывания, а еще точнее — ее зависимость от условий, в которых происходит реакция.

Описание некоторых других кодов можно найти в [8]. Теперь в п. 3—6 мы более подробно опишем несколько фрагментов разных кодов из разных генетических языков. К общим проблемам мы вернемся в п. 7.

3. Первый пример. Регуляция трансляции рибосомальных белков у прокариот. Как уже было сказано, трансляция осуществляется рибосомами. Все рибосомы в клетке одинаковы и каждая состоит из нескольких рРНК и нескольких десятков рибосомальных белков, каждый из которых имеет свое собственное уникальное место в структуре рибосомы. Поскольку рибосом клетке требуется очень много, все время происходит интенсивный синтез рРНК и белков и сборка новых рибосом. При этом возникает необходимость поддержания правильного соотношения между синтезом различных бел-

ков с тем, чтобы не допустить перепроизводства отдельных белков. Эта регуляция осуществляется следующим образом.

Гены рибосомальных белков собраны в несколько оперонов. В регуляторной зоне каждой мРНК имеется структурная особенность, узнаваемая одним из белков, закодированных на данной мРНК. Связываясь с этим участком, белок подавляет трансляцию всей матрицы. Поэтому, если при сборке остаются лишние молекулы белков из этого оперона, и, в частности, лишние молекулы регулирующего белка, то они связываются с матрицами и временно прекращают трансляцию всех этих белков. Самое интересное — это то, что каждый регуляторный участок представляет из себя элемент вторичной структуры — шпильку, подобную той, которая имеется на рибосомальной РНК и узнается данным белком при сборке рибосом (рис. 1). Этот пример показывает, что невозможно провести четкую грань между функциональными и структурными кодами (в этом случае — кодами белок-РНКового взаимодействия).

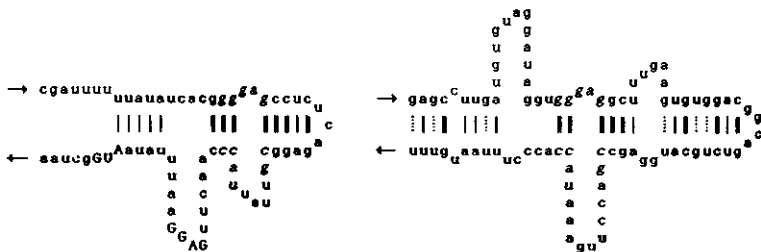


Рис. 1. Сравнение вторичных структур РНК, узнаваемых рибосомальным белком L1 на мРНК при подавлении трансляции (слева) и на рРНК при сборке рибосомы (справа). Вертикальными черточками показано спаривание комплементарных и взаимодействие некоторых некомплементарных нуклеотидов. Курсивом показаны одинаковые участки узнавания на обеих РНК. Большими буквами на мРНК выделены стартовый кодон AUG и участок GAGG, узнаваемый рибосомой перед началом трансляции.

4. Второй пример. Регуляция транскрипции оперонов аминокислотного синтеза у проکاریот. Кишечная палочка способна синтезировать все необходимые аминокислоты, используя в качестве субстрата сахар, например, глюкозу или лактозу. Отдельные этапы аминокислотного синтеза регулируются белковыми ферментами, а гены белков, отвечающих за отдельные этапы синтеза какой-либо аминокислоты, собраны в опероны. Ясно, что каждый такой оперон должен транскрибироваться только при дефиците соответствующей аминокислоты. В этом пункте мы рассмотрим, как осуществляется эта регуляция.

Типичная регуляторная область аминокислотного оперона показана на рис. 2. Она состоит из промотора, участка, кодирующего короткий лидерный белок, и частично перекрывающихся с ним четырех участков А, Б, В, Г, которые могут образовывать три альтернативные вторичные структуры АБ, БВ и ВГ, и, наконец, последовательность из нескольких Т подряд. При этом в лидерном гене содержится много кодонов аминокислоты, синтез которой катализируется белками данного оперона, шпилька (элемент вторичной структуры) АБ кодирует паузу — приостановку транскрипции, шпилька ВГ — это часть терминатора транскрипции, а вторая часть этого терминатора — это упомянутая последовательность из нескольких Т (точнее, полимераза прекращает транскрипцию, когда на транскрипте образуется шпилька и за ней — несколько U; предполагается, что шпилька вызывает приостановку транскрипции, а ряд U — высвобождение транскрипта).

Узнав промотор, РНК-полимераза начинает транскрипцию и останавливается, синтезируя участок А и Б, которые образуют шпильку АБ. Во время паузы рибосома узнает сигнал начала трансляции, расположенный на уже синтезированной части РНК, и начинает синтез лидерного белка. Дойдя до шпильки АБ, рибосома освобождает РНК-полимеразу, которая возобновляет транскрипцию. Далее, в зависимости от концентрации данной аминокислоты, на транскрипте образуется одна из альтернативных шпилек БВ и ВГ (напомним, что шпилька АБ уже расплетена рибосомой). Если аминокислоты достаточно, рибосома следует за РНК-полимеразой до стоп-кодона, после чего покидает транскрипт. На РНК же образуется шпилька ВГ (рибосомы уже нет,

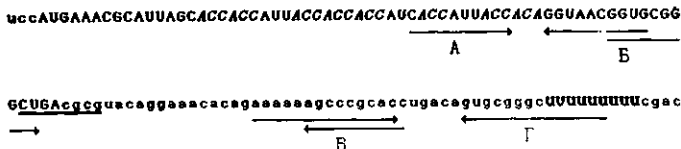


Рис. 2. Фрагмент регуляторной области в транскрипте оперона синтеза аминокислоты треонина кишечной палочки. Большими буквами показана область, кодирующая лидерный пептид, в котором курсивом выделены треониновые кодоны ACC и ACA. Стрелками показаны взаимно-комплементарные участки. Подчеркнута область, в которой происходит приостановка транскрипции при образовании шпильки АБ. Жирным шрифтом выделена последовательность U...U, которая вместе со шпилькой ВГ образует терминатор транскрипции.

и она не мешает), которая вместе со следующей за ней последовательностью поли-У вызывает прекращение транскрипции. Таким образом, основная часть оперона остается нетранскрибированной. Если же клетка испытывает аминокислотное голодание, рибосома синтезирует лидерный белок медленно, застревая на кодонах данной аминокислоты в ожидании соответствующих тРНК, поэтому она не успевает за РНК-полимеразой, и образуется шпилька БВ. Тем самым участок В уже занят, и терминатор, необходимой частью которого является шпилька ВГ, отсутствует. Поэтому РНК-полимераза продолжает транскрипцию и синтезирует длинную мРНК.

5. Третий пример. Регуляция транскрипции белковых генов у высших эукариот. Основная регуляция белкового синтеза у эукариот осуществляется на уровне транскрипции. Это связано с тем, что белки эукариот существенно (в сотни раз) разнообразнее, многие из белков нужны лишь в определенных тканях, а мРНК стабильнее и затраты на их производство больше. Поэтому синтезировать все возможные мРНК было бы неоправданным расточительством и основная регуляция белкового синтеза осуществляется на уровне транскрипции. Соответственно, белковые факторы, определяющие наличие и уровень транскрипции, существенно более разнообразны, а регуляторные области устроены более гибко и разнообразно, чем у прокариот.

Кроме промоторов — участков, непосредственно примыкающих к точке старта транскрипции, в ДНК эукариот имеется второй тип регуляторных областей — энхансеры. Эти области влияют на уровень транскрипции, находясь на расстоянии нескольких тысяч нуклеотидных пар от точки старта, причем уровень транскрипции практически не зависит от последовательности между энхансером и промотором и слабо зависит от расстояния, которое может меняться в широких пределах как в эксперименте, так и в самой клетке.

Обязательным элементом промотора является сигнал (так называемый ТАТА-бокс или его эквивалент), находящийся на фиксированном расстоянии от точки старта транскрипции. При отсутствии этого сигнала транскрипция не происходит. В то же время связывания соответствующего фактора с этим сигналом обычно бывает недостаточно и для инициации транскрипции необходимо связывание еще каких-либо факторов с соответствующими участками ДНК, причем различные факторы могут узнавать один и тот же участок ДНК, что создает дополнительные возможности для регуляции. По-видимому, РНК-полимераза узнает не какой-то сигнал, а набор белковых факторов, связанных с ДНК (рис. 3). Расстояние между точкой старта транскрипции и сайтами связывания дополнительных общих и тканеспецифичных факторов бывает равно 30—110 нуклеотидных пар и жестко не фиксировано.

Энхансер состоит из нескольких модулей, находящихся на расстоянии в 50—100 нуклеотидных пар друг от друга, а каждый модуль в свою очередь составлен из одного или нескольких энхансонов — сайтов связывания отдельных белков. В то время, как модули могут быть довольно свободно переставлены и скомбинированы с приданием энхансеру новых свойств (например, другой силы и/или тканевой специфичности), узнавание энхансонов соответствующими белками, по-видимому, часто происходит кооперативно и поэтому отдельные энхансоны прочнее связаны друг с другом, возможности для перекombинации ограничены, а расстояние между ними фиксировано гораздо более жестко.

6. Четвертый пример. Генетический переключатель фага λ. Это регуляторная область, состоящая из двух расположенных на комплементарных цепях промоторов

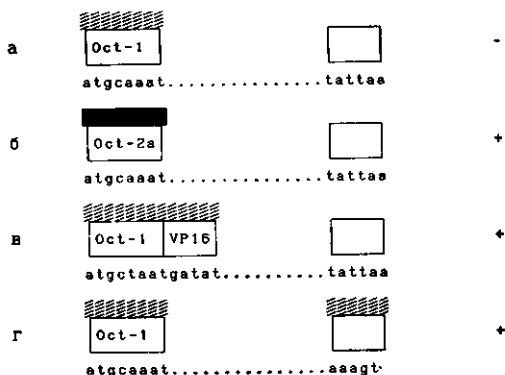


Рис. 3. Пример регуляции транскрипции у эукариот. Штриховкой показано средство белковых факторов к РНК-полимеразе.

а) В большинстве клеток связывания общего фактора с TATA-боксом *tattaa* и Oct-1 с октамером *atgcaaat* недостаточно для инициации транскрипции.

б) В лимфоцитах с октамером связывается специфический фактор Oct-2a, обладающий сильным средством к РНК-полимеразе, и тот же промотор иницирует транскрипцию.

в) Белок-активатор VP16 вируса герпеса взаимодействует с Oct-1 и этот комплекс, обладающий достаточным средством к РНК-полимеразе, узнает расширенный октамер в вирусных промоторах даже в нелимфоидных клетках.

г) Другой промотор содержит эквивалент TATA-бокса *aaagt*, который узнается фактором, имеющим собственное средство к РНК-полимеразе; в результате полимераз взаимодействует с этим фактором и Oct-1.

P_R и P_{RM} , инициирующих транскрипцию мРНК белков *cI* и *Cro* соответственно (само переключение состоит в альтернативном синтезе этих белков), и трех перекрывающихся с ними симметричных операторов O_{R1} , O_{R2} и O_{R3} , с которыми могут связываться эти белки (рис. 4). Связывание любого белка с O_{R1} подавляет транскрипцию гена *cro* (начинающуюся с промотора P_R), а с O_{R3} — транскрипцию гена *cI* (промотор P_{RM}). Связываясь с O_{R2} , белок *Cro* подавляет транскрипцию своего гена *cro*, а *cI* — тоже выключает P_R , но, кроме того, стимулирует P_{RM} . В отличие от P_R , этот промотор слабый и без дополнительной стимуляции (происходящей при взаимодействии РНК-полимеразы с белком *cI*, сидящем на O_{R2}) он не может иницировать транскрипцию. Последний факт, который необходим перед описанием самого механизма переключения, состоит в том, что *cI* предпочтительно связывается с O_{R1} , что облегчает связывание его с соседним O_{R2} (при отсутствии кооперативности средство *cI* к O_{R2} и O_{R3} одинаково), а связывание *Cro* некооперативно и происходит сначала с O_{R3} .

Итак, нам необходимо осуществлять выбор между синтезом *cI* и *Cro* (это решение в дальнейшем запустит другие регуляторные механизмы, на которых мы не будем здесь останавливаться; они описаны в [9]). В начальном состоянии работает ген *cI*, и молекулы *cI* связаны с O_{R1} и O_{R2} . Если происходит перепроизводство, то лишние молекулы *cI* связываются также и с O_{R3} и подавляют транскрипцию своего гена (следует иметь в виду, что связывание белков с ДНК — динамический процесс и точнее было бы говорить о среднем времени, которое молекулы белка проводят на соответствующих сайтах). При облучении клетки-хозяина ультрафиолетом (в этот момент фагу необходимо совершить переключение с синтеза *cI* на синтез *Cro*) один из ее белков модифицирует молекулы *cI* таким образом, что они не способны более к кооперативному узнаванию операторов. Поэтому *cI* освобождает сначала O_{R2} (и тем

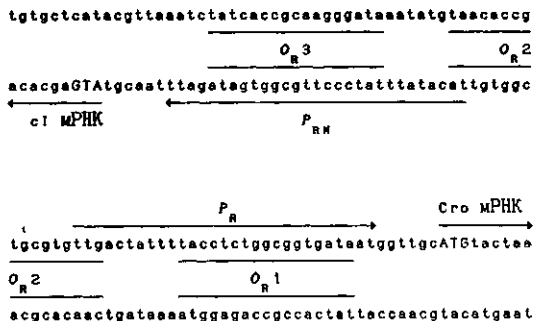


Рис. 4. Генетический переключатель фага λ . Показаны обе цепи ДНК, на которых выделены функциональные участки: промоторы P_R и P_{RM} (с задаваемыми ими направлениями транскрипции), симметричные операторы $O_R 1$, $O_R 2$ и $O_R 3$, а также начала двух мРНК, большими буквами на которых показаны стартовые кодоны ATG).

самым прекращается транскрипция гена cI с промотора P_{RM} , а затем и $O_R 1$. После освобождения $O_R 1$ начинается синтез белка Cro (напомним, что промотор P_R гена cro не нуждается в стимуляции), который связывается сначала с $O_R 3$, окончательно подавляя синтез cI , а затем, по мере повышения концентрации, и с $O_R 2$ и $O_R 1$, в конце концов подавляя транскрипцию своего гена с промотора P_R . Переключение произошло (на самом деле промотор P_R инициирует транскрипцию целого оперона, в который входит ген cro ; дальнейшая регуляция осуществляется другими новосинтезированными белками этого оперона).

7. Свойства генетического языка. Итак, в генетическом языке сложным образом сосуществует множество кодов, фрагменты которых собраны в отдельные подсистемы. Ситуация осложняется тем, что для большинства кодов нам не известны их элементарные единицы; более того, неразложимые единицы одного кода могут тем не менее находиться в простом соотношении с составными единицами другого кода, например — аминокислоты и кодоны, связанные посредством генетического кода. Возможно, что нечто подобное имеет место для элементов пространственной структуры.

Мы не знаем, что является единицами низшего уровня, например, в кодах взаимодействия белков и ДНК или белков и РНК (следует отметить, что здесь было бы точнее говорить о взаимно согласованных кодах — белковом и нуклеиновом). Как уже отмечалось выше, простейшая аналогия фонема — нуклеотид (или фонема — аминокислота) кажется нам не вполне правильной, точнее недостаточно полной, потому что кроме непосредственного узнавания конкретного нуклеотида в конкретной позиции, для белка важна пространственная конфигурация участка, элементы которой зависят от нескольких соседних нуклеотидов, в то время как каждый нуклеотид может влиять на несколько элементов пространственной структуры.

Поднимаясь на следующий уровень, мы встречаем единицы, имеющие значение (т. е. способность к взаимодействию с другими единицами), но функционально не самостоятельные. Сюда можно отнести энхансоны, другие сайты кооперативного связывания отдельных белков, отдельные элементы вторичной структуры (например, шпильки в терминаторе тран-

скрипции прокариот) и т. д. Сочетаемость этих элементов довольно ограничена. Направляется соотношение этого уровня с морфологическим уровнем естественного языка.

Еще выше располагаются комбинации элементов морфологического уровня (промоторы, терминаторы, энхансерные модули и т. д.). Их сочетаемость существенно шире, а различные их комбинации обеспечивают возможность гибкой регуляции. Эти единицы можно считать аналогом слов. Сюда же относятся сайты некооперативного связывания белков («слова», состоящие из единственной морфемы).

Наконец, набор регуляторных областей вместе с регулируемыми генами является совершенно самостоятельной единицей (предложением), функционирование которой не зависит от положения в геноме. Набор этих предложений — молекула ДНК — является текстом генетического языка.

Заметим, что мы не задавались изначальной целью найти аналог фонемам, морфемам, словам и предложениям, а рассматривали лишь иерархию молекулярно-генетических структур, их значения и дистрибуцию. Выделение этих уровней произошло в некотором смысле само собой.

Как и естественный язык, большинство кодов генетического языка выродены (одно сообщение может быть закодировано многими способами), и это дает возможность одновременного кодирования нескольких сообщений на одном фрагменте ДНК. С другой стороны, многие участки, в особенности у эукариот, по-видимому, не несут решительно никакой смысловой нагрузки.

Так же как и в естественном языке, в генетическом языке имеется два вида синонимии. Синонимия первого вида возникает, когда в какой-либо ситуации несущественно различие между двумя элементами. Примером могут быть синонимичные кодоны, кодирующие одну и ту же аминокислоту или, скажем, два различных промотора транскрипции. Заметим, что во втором примере мы встречаемся с общесемiotическим противопоставлением знака (sign) и экземпляра знака (token). Биологическая терминология здесь не устоялась, и слово «промотор» может означать в зависимости от контекста любую область инициации транскрипции, узнаваемую специфическим набором белков, и конкретную последовательность перед данным опероном. Синонимия же второго вида — это именно синонимия различных (наборов) знаков; например, синонимия промоторов, описанная в п. 4 и показанная на рис. 3.

8. В заключение еще раз следует подчеркнуть, что наши представления о молекулярно-генетических регуляторных системах пока еще крайне неполны и отрывочны. Все время появляются новые факты, разрушающие казалось бы самые устоявшиеся догмы. Несмотря на это, мы уже можем приблизительно сформулировать хотя бы некоторые самые общие принципы построения генетического языка. Целью настоящей работы было показать, что многие из этих принципов находят аналог в естественном языке *.

* Автор выражает благодарность С. А. Крылову, Е. В. Кузину, С. А. Старостину и Ю. С. Степанову, беседы с которыми способствовали прояснению многих высказанных утверждений; разумеется, ответственность за оставшиеся неясности целиком лежит на авторе статьи.

СПИСОК ЛИТЕРАТУРЫ

1. Jakobson R. Life and language // Linguistics. 1974. № 138.
2. Jakob F. The linguistic model in biology // Roman Jakobson. Echoes of his scholarship / Ed. by Armstrong D., Schooneveld C. H. van. Lisse, 1977.
3. Гамкрелидзе Т. В. Р. О. Якобсон и проблема изоморфизма между генетическим кодом и семиотическими системами // ВЯ. 1988. № 3. С. 6—7.
4. Smith T. F. Semantic and syntactic patterns in the genetic language // Biomolecular data: A resource in transition / Ed. by Swartz D. G., MacDonell M. T., Colwell R. R., Oxford; New York, 1989.
5. Ратнер В. А. Молекулярно-генетические системы управления. Новосибирск, 1975.
6. В мире науки. 1985. № 12 («Молекулы жизни»).
7. Уотсон Дж. Д. Молекулярная биология гена. М., 1978.
8. Trifonov E. N. Codes of nucleotide sequences // Mathematical biosciences. 1988. V. 90. P. 507.
9. Пташине М. Переключение генов. М., 1988.