КОМПЬЮТЕРНОЕ МОДЕЛИРОВАНИЕ В ЗАДАЧАХ РЕГУЛЯЦИИ РАБОТЫ ГЕНОВ И ЭВОЛЮЦИИ ОРГАНИЗМОВ

В.А. Любецкий
Институт проблем передачи информации имени А.А. Харкевича РАН

Кроме сведений общекультурного характера в области молекулярной биологии заметка содержит краткие формулировки нескольких проблем в области математической биологии, биоинформатики. Заметка бегло знакомит читателя с тематикой одной из лабораторий Института проблем передачи информации РАН.

Ключевые слова: *ген, регуляторная система, эволюция, кластеризация белков, взаимодействие и конкуренция РНК-полимераз, согласование деревьев генов и видов, аттенюаторная регуляция транскрипции, трехмерная и одномерная диффузия в клетке, происхождение видов.*

COMPUTER MODELING IN PROBLEMS OF GENE REGULATION AND SPECIES EVOLUTION

V.A. Lyubetsky

This is a concise review on the problems definition that also sketches the activity of one of research groups in the Institute for Information Transmission Problems. Apart from giving a common general background, it introduces the reader into particular definitions. Therefore, references include both textbooks and only selected publications of the group.

Keywords: *gene, regulatory system, evolution, protein clustering, interaction and competition of RNA-polymerases, reconciliation of gene and species trees, transcription attenuation, spatial and linear diffusion in cell, origin of species.*

Сравнительно недавно Высшая аттестационная комиссия при Министерстве образования и науки РФ открыла специальность 03.01.09 «Математическая биология, биоинформатика», по которой можно получить степени в области физико-математических, биологических и медицинских наук; а также родственную специальность 03.01.08 «Биоинженерия». В нашей стране по первой из этих специальностей работают шесть диссертационных советов: в Институте проблем передачи информации им. А.А. Харкевича РАН, Сургутском государственном университете, Российском государственном медицинском университете, Юго-Западном государственном университете, Научно-исследовательском институте биомедицинской химии им. В.Н. Ореховича РАМН, Институте цитологии и генетики СО РАН.

В мире исследования в области математической биологии и биоинформатики занимают огромное место, прежде всего, в связи с обширным спектром приложений, непосредственно связанных с человеком и его текущей жизнедеятельностью. Это, в частности, — медицина, фармакология, парфюмерия, пищевая промышленность, ветеринария, очистка среды от любых загрязнений (тяжелыми металлами, радиоактивными изотопами и т.д.).

Трудно переоценить и фундаментальную роль этих исследований. Общая задача естественных наук состоит в точном описании происходящих в природе процессов, в создании теоретических моделей этих процессов. Такое описание по общепринятому мнению должно быть «точно сформулированным», математическим. Биология, которая стоит за этой специальностью, в значительной мере, — молекулярная биология, биология процессов, происходящих в клетке или между клетками, процессов, связанных с рождением, распадом, преобразованием, воздействием друг на друга биологических молекул (прежде всего, нуклеиновых кислот и белков). Иными словами, мы говорим о генетике, которая обращается к вопросам: как из нескольких молекул ДНК развивается жизнь, как на их основе функционирует клетка и организм, как возникла и изменяется сама молекула ДНК. Это поднимает нас к вопросам возникновения жизни, психики, речи и сознания, включая социальную организацию людей и животных. Если к этим вопросам добавить еще вопрос о возникновении и развитии Вселенной, то казалось бы можно сказать, что все они находятся на недоступном нам уровне. Однако в последние десятилетия некоторые из них, в том числе, вопросы функционирования клетки и

эволюции организмов прояснились в значительной степени.

В последние десятилетия появились огромные базы данных геномной информации, одна из самых известных — GenBank.

Молекула ДНК, по современным представлениям полный источник жизни, это — просто последовательность в 4-буквенном алфавите с типичной длиной около 3 млн букв у одноклеточных организмов (бактерий) и 3 млрд букв у многоклеточных организмов (животных). В упомянутой базе данных собрано огромное количество таких последовательностей. Их можно сравнивать между собой и, тем самым, извлекать новое биологическое знание не из эксперимента в традиционном смысле (как говорят, «мокрого опыта», «опыта в пробирке»), а из «компьютерного опыта». Компьютерный сравнительный анализ последовательностей из GenBank и других баз данных уже привел ко многим биологическим открытиям в области функционирования и эволюции клетки.

В последние годы сделан следующий шаг. Появились математические и компьютерные модели функционирования клетки и эволюции организма. Эти модели носят механистический характер, так как подлинная физика поведения биологических молекул описывается по современным представлениям квантовой теорией, слишком сложной для получения реальных решений. Однако и эти модели в основном не допускают пока строго математического решения, их исследование, предсказание на их основе биологических феноменов осуществляются с помощью моделирования, которое также требует многих часов и иногда многих суток работы суперкомпьютера. Мы обычно используем кластер MVS-100K в МСЦ РАН.

Тем не менее, поразительно, что такие модели, раскрывая внутреннюю механику функционирования клетки и эволюции организма, предсказывают результаты, которые сходятся с известными из мокрых опытов результатами с той точностью, которая доступна в самой мокрой лаборатории. Более того, эти модели предсказывают новые биологические явления молекулярного уровня, которые были проверены в лаборатории и подтвердились после того, как были получены на компьютере.

Различие слов «Математическая биология» и «Биоинформатика» можно понимать таким образом. Слова «Математическая биология» относятся к науке, подобной теоретической физике, которая разрабатывает именно математические и компьютерные модели явлений, объясняющие опыты и предсказывающие новые значения параметров для их проверки в последующих «мокрых» опытах, как это происходит в теоретической физике. Слово «Биоинформатика» акцентирует внимание на компьютерном анализе и непосредственном сравнении данных, включая создание и расширение самих баз данных и их функций (методов поиска информации в них). Конечно, оба эти подхода дополняют друг друга и в реальном исследовании работают вместе: сравнение данных уже предполагает некоторую, может быть, простую модель явления, а создание более сложной модели опирается на предварительные результаты сравнения и требует самих этих данных для реального счета.

Лаборатория «Математических методов и моделей в биоинформатике» Института проблем передачи информации РАН (ИППИ РАН) занимается проблемами регуляции работы генов и эволюции регуляторных систем, генов и видов. Лаборатория имеет около 400 публикаций, включая тезисы международных конференций, и ведет постоянный семинар-курс по этой тематике на механико-математическом факультете МГУ.

Прежде чем перейдем к обсуждению ряда конкретных проблем, связанных с регуляцией работы генов в живой клетке и с эволюцией организмов, начнем с краткого общекультурного введения.

Геном — набор небольшого (десятки) числа молекул ДНК, т.е. разных последовательностей (суммарной длиной миллионы или миллиарды) букв в 4-х буквенном алфавите {A, T, G, C}, а ген — направленный участок в одной из этих последовательностей, т.е. также последовательность букв, но гораздо более короткая. Последовательности, составляющие геном, и, в частности, гены являются молекулами нуклеиновой кислоты. Число генов может быть до 8–9 тысяч (у бактерий) и до 50 тысяч (у животных). Клетка имеет не более трех мест хранения этих последовательностей и их генов. Эти места называются ядром, митохондрией и пластидой (а в целом: органеллами клетки); все три органеллы присутствуют, например, у растений, только первые две — у животных. Соответствующие гены называются ядерными, митохондриальными (митохондриомными) и пластидными (пластомными). Число генов может быть и малым: несколько сот (у бактерий, в пластидах) или несколько штук (в митохондриях). Между генами расположены участки, которые регулируют активность (интенсивность работы) генов. Как и все в геноме, они являются направленными последовательностями (обычно, более короткими, чем гены) букв и кодируют сложные структуры. Таким образом, геном — это множество генов и регуляторных участков («регуляторных систем»), расположенных не более, чем в трех органеллах клетки. Кроме них в геноме имеются обширные участки, назначение которых неизвестно, их можно назвать «бессмысленными»; они не получают теоретического описания и не участвуют в моделировании. Во многих задачах буквенный состав генома, гена, регуляторной системы не так важен.

Некоторые гены кодируют последовательности, уже не являющиеся нуклеиновыми кислотами, это — последовательности не в 4-х буквенном алфавите,

а в 20-ти буквенном алфавите, которые называются белками. Длина белка равна примерно третьей части длины гена, кодирующего этот белок. Нуклеиновые кислоты и белки существенно различаются по многим свойствам. Функционирование клетки основано на том, что некоторые гены чрезвычайно сложным образом преобразуются в белки, а полученные белки регулируют ход этого преобразования. Таким образом, мы видим здесь типичное для математики и информатики явление рефлексии (обращения внимания процесса на самого себя).

Организм в интересующем нас аспекте — это геном (в его прижизненном развитии). Вид — это совокупность организмов (геномов) с близкими характеристиками, что позволяет виду иметь потомство, которое также состоит из организмов (геномов). Таким образом, развитие вида от самого исходного, предкового генома состоит в последовательном рождении/гибели и преобразовании геномов.

Все процессы развиваются в физическом времени, хотя до сих пор в моделях принято для упрощения (возможно, кажущегося) рассматривать дискретное время, которое в математическом смысле представляется графом, а обычно — деревом.

В заключение этого общекультурного введения назовем на более профессиональном языке список проблем, которыми занимается лаборатория в части указанной специальности:

1) поиск регуляции (сайтов, факторов, вторичных структур) экспрессии генов на уровнях транскрипции, трансляции и процессинга; выяснение механизмов регуляций и исследование эволюции регуляций у бактерий, растений, водорослей, простейших, беспозвоночных, хордовых и т.д., их пластид и митохондрий; построение функционально сходных семейств белков (органелл и др.), специфичных для узких таксономических групп; разработка баз данных биоинформации;
2) совместная эволюция регуляторных систем, генов и видов и согласование этих эволюций между собой: от бактерий и архей до млекопитающих; согласование деревьев генов и видов, согласование набора деревьев генов (построение супердерева), построение дерева генов по множественному выравниванию последовательностей;
3) роль горизонтальных переносов генов и других эволюционных событий в эволюции (построение эволюционных сценариев); происхождение видов из разрозненных геномов.

Буквально в последние несколько лет появились точно сформулированные описания биологических процессов молекулярного уровня. Хотя в них используется несложная математика, до сих пор не удается провести строгого исследования ни одной из возникающих задач, что было бы очень желательно для получения биологических выводов. Математическое исследование заменяется компьютерным моделированием, которое в этой области приобрело нетривиальный характер также только в последние годы. Мы приведем несколько примеров процессов, которые, безусловно, содержательно описывают биологические явления.

Задачи 1–5 на уровне моделирования исследованы нами. Для них были построены численные компьютерные модели, которые дали результаты, близкие к экспериментальным данным, обычно даже находящиеся в пределах экспериментальных ошибок. Напротив, для задач 6–7 не ясно даже, как эффективно проводить моделирование. Не всегда доступны прямые экспериментальные измерения величин, о которых говорится в задачах 1–7. Иногда приходится сравнивать модель с косвенными биологическими данными, что, однако, является обычной ситуацией в естественных науках.

Биологические термины, упоминаемые и отчасти поясняемые ниже, не существенны для математического понимания дальнейшего.

ПРОБЛЕМЫ И НЕКОТОРЫЕ РЕЗУЛЬТАТЫ МОДЕЛИРОВАНИЯ

1. Классификация белков. Дано множество всех пластомных белков, например, из водорослей или, более обычно, из группы родственных водорослей. Найти кластеризацию (т.е. разбиение этого множества белков на попарно не пересекающиеся подмножества), так чтобы в один кластер попали «родственные» белки и только они. Образно говоря, «родственными» называют «один и тот же» белок, «две копии» одного белка, находящиеся в пластидах одного или разных видов. Эти кавычки поясняют еще так: эти белки «мало отличаются», как две последовательности, выполняют «одну и ту же функцию» в клетке, имеют «общее происхождение» от некоторого предкового белка. Такие белки называют гомологичными (или ортологичными). Итак, задача состоит в нахождении кластеров (семейств) гомологичных белков. Разумеется, кластеризация и различные алгоритмы для ее выполнения давно и широко используются.

Нами предложен следующий простой алгоритм для нахождения упомянутых кластеров, результаты применения которого хорошо согласуются с биологическими наблюдениями. Математически говоря, набор семейств гомологичных белков можно определить как результат работы этого алгоритма; и это — типичная цель любой модели.

Эффективная компьютерная реализация для суперкомпьютера этого и других упоминаемых ниже алгоритмов остается актуальной задачей.

Итак, пусть задан набор пластид, которые индексируем буквой i, и для каждой пластиды заданы ее белки P_{ij}, индексируемые буквой j. Для всех пар белков (P_{ij}, P_{kl}) из всех пар пластид вычисляется характеристика близости $s_0(P_{ij}, P_{kl})$ белков; известны алгоритмы,

которые это делают, и мы их здесь не обсуждаем. Наш алгоритм вычисляет нормированную *степень сходства* $s(P_{ij}, P_{kl})$ белков $2s_0(P_{ij}, P_{kl})(s_0(P_{ij}, P_{kl})+s_0(P_{ij}, P_{kl}))^{-1}$.

Затем рассматривается *полный* неориентированный граф G с множеством вершин $\{P_{ij}\}$, в котором каждому ребру (P_{ij}, P_{kl}) приписано значение $s(P_{ij}, P_{kl})$, которое будем называть *весом* этого ребра; петли отсутствуют. Для уменьшения объема вычислений вместо полного графа можно использовать *разреженный* граф G из ребер (P_{ij}, P_{kl}), удовлетворяющих условиям: $s(P_{ij}, P_{kl})=\max_m s(P_{ij}, P_{kl})=\max_m s\,(P_{ij}, P_{km})$, $s(P_{ij}, P_{kl})\geq L$, где максимумы берутся по всем белкам из соответствующих пластид i-й и k-й, а L — параметр алгоритма. Если $i=k$, то предполагается еще условие: $m\neq l$ и второе равенство нужно опустить. Для полученного графа G алгоритм строит набор всех его связных компонент. Затем для каждой компоненты (обозначим ее той же буквой G) строятся «накрывающие» G бескорневые деревья D. Последнее означает: в G перебираются ребра в порядке убывания их веса, которые объявляются ребрами строящегося дерева D; если добавление к D очередного ребра из G приводит к появлению в D цикла, то это ребро пропускается. В результате D не содержит циклов, т.е. является деревом, и включает все вершины из G. Сумма весов всех ребер в D называется *весом* дерева D; полученные деревья имеют максимально возможный вес. Итак, для каждой связной компоненты в исходном графе G строятся накрывающие ее деревья D. Затем к каждому дереву D применяется следующая рекурсивная процедура «разделения дерева», которая строит набор деревьев $\{D_{ij..}\}$, в котором все индексы равны 1 или 2. Длина последовательности $ij..$ индексов называется *глубиной* дерева $D_{ij..}$. Если текущее дерево, скажем, $D_{ij..}$ некоторой глубины k из этого набора не удовлетворяет сформулированному ниже критерию разделения дерева, то оно заменяется в наборе на два новых дерева $D_{ij..1}$ и $D_{ij..2}$ глубины $k+1$ каждое путем удаления из $D_{ij..}$ ребра e_0 с минимальным по всему $D_{ij..}$ весом s, если при этом выполнено $s<H$, где H — параметр алгоритма (если это неравенство не выполняется, то текущее дерево не делят, и переходят к следующему дереву). Иначе проверяем критерий сохранения текущего дерева $D_{ij..}$ без изменения. Этот критерий для дерева $D_{ij..}$ с множеством вершин V состоит в выполнении трех условий: (1) $|V|\leq pn$, где $|V|$ — число вершин в дереве $D_{ij..}$, а n — число всех пластид в исходном наборе и p — параметр алгоритма; (2) ребро (P_{mq}, P_{kl}) с минимальным весом в $D_{ij..}$ соединяет белки P_{mq} и P_{kl}, у которых $m\neq k$; (3) любая пара вершин P_{mq} и P_{ml} дерева $D_{ij..}$, соответствующих белкам из одной пластиды, соединена в $D_{ij..}$ путем, состоящим из вершин, соответствующих белкам той же пластиды. Если этот критерий выполнен и имеется следующее дерево, то переходим к нему. Если все деревья уже исчерпаны, то алгоритм завершает работу. Полученный в результате набор деревьев представляет собой разбиение исходных белков на кластеры, состоящие из последовательностей, приписанных всем вершинам одного дерева.

2. Конкурирующие процессы связывания и движения (конкуренция РНК-полимераз). Дана последовательность в четырехбуквенном алфавите, на которой отмечены направленные участки двух типов: одни называются генами, другие промоторами. Геометрия расположения генов и промоторов может быть произвольной, но она фиксирована. С каждым промотором, если он свободен, связывается молекулярная машина («полимераза») одного из фиксированного конечного числа типов. Полимераза имеет фиксированную длину и движется по направлению промотора, вообще говоря, вдоль всей последовательности. Таким образом, много разных полимераз одновременно связываются с последовательностью и движутся по ней, каждая в своем направлении (из двух возможных). Промотор «свободен» в данный момент, если в его пределах не находится никакой части никакой полимеразы. Ген «считывается», если некоторая полимераза прошла по его направлению от его начала до его конца. Частота считывания гена называется его «уровнем транскрипции». Каждый промотор для каждого типа полимераз характеризуется своей интенсивностью *попыток связывания* с ним полимераз этого типа. Можно считать, что концентрация любого типа полимераз достаточная, т.е. интенсивность отражает только качество самого промотора для данного типа полимераз. Попытка считается осуществленной, если промотор свободен в момент ее реализации. Для *части типов* после связывания происходит «абортивный» процесс, состоящий в чередовании движения с конечной скоростью по направлению промотора на случайное расстояние и в мгновенном возвращении в исходное положение. Такие односторонние колебания продолжаются случайное число раз до тех пор, пока полимераза не отойдет на критическое расстояние от промотора. В этот момент полимераза *отрывается* от промотора и ее длина мгновенно уменьшается на известную величину, и движение в том же направлении продолжается. Для *оставшихся типов* абортивный процесс отсутствует, движение начинается сразу после связывания, длина полимеразы не меняется. Допустимо, что попытки образуют пуассоновский процесс, а полимераза движется детерминировано с фиксированной скоростью, своей для каждого типа, вплоть до столкновения с другой полимеразой. После столкновения двух полимераз, движущихся друг за другом в одном направлении, скорость первой не меняется, а скорость второй ограничивается скоростью первой до тех пор, пока первая связана с последовательностью («элонгирует»). В случае встречного движения обе полимеразы покидают последовательность («терминируют»). Здесь биологический интерес представляют многие задачи,

например: даны интенсивности попыток связывания всех промоторов, найти уровни транскрипции всех генов. Обратная задача: даны уровни транскрипции всех генов, найти интенсивности попыток связывания, которые приводят к наилучшему приближению этих уровней. Еще задача: даны простые законы изменения во времени уровней транскрипции генов и скоростей всех полимераз (в ситуации значительного изменения температуры), найти в том же смысле интенсивности попыток связывания. Большие проблемы возникают, если отказаться от предположения о детерминированном характере движения полимераз, что биологически более адекватно. Стохастическое движение полимераз строго описано нами, но получается слишком сложная задача даже для моделирования. Предложенное нами компьютерное решение доступно по адресу. Простой случай, хотя биологически мало интересный, возникает, если предположить отсутствие абортивного процесса, равенство между собой всех скоростей полимераз и нулевые размеры всех промоторов и полимераз. Тогда задача сводится к специальному случаю теории встречных потоков с аннигиляцией. Дополнительные трудности возникают, если вместо последовательности рассматривается замкнутая («кольцевая») последовательность (т.е. ее буквы как бы равномерно расположены по некоторой окружности). Например, конкуренция на кольцевой последовательности с длиной в 17 тысяч букв (митохондриальный геном человека). В нем присутствуют полимеразы только одного типа, а три промотора расположены вблизи следующих позиций: 407 против часовой стрелки, 561 и 646 по часовой стрелке. Абортивные процессы отсутствуют. Сначала полимеразы не проходят полный круг, встречные потоки полимераз с трех промоторов сталкиваются и срываются. Поэтому дальние от промоторов гены имеют почти нулевые уровни транскрипции, что не соответствует биологической реальности. Это состояние кажется неустойчивым: в какой-то момент число связываний с одним из промоторов оказывается больше (на 10–20 полимераз), эти «лишние» полимеразы не аннигилируют, проходят полный круг и, в том числе, свой промотор. Последнее создает эффект роста интенсивности связывания с этого промотора, благодаря чему происходит рост числа «круговых» полимераз в одном направлении. Если случайно в достаточной мере возрастет число связываний с другого промотора, то направление процесса может поменяться. Направление редко меняется несколько раз. Быстро устанавливается преобладающее направление потока полимераз. Как только число «круговых» полимераз в одном направлении превзойдет некоторый порог, интенсивность эффективного связывания с одним из промоторов и уровень транскрипции соответствующих генов будут постоянно увеличиваться. И так вплоть до заполнения полимеразами всей последовательности (с промежутками менее чем длина полимеразы). Задача: описать эти режимы и бифуркации. Обычно на окружности в определенных местах имеются еще «протекающие терминаторы». Это — сайты, которые в каждом из направлений пропускают только свою в среднем фиксированную долю полимераз. При мутациях, разрушающих эти сайты, возникают тяжелые заболевания. Какова динамика процесса в этом случае? Здесь геометрия расположения может быть также весьма разной. Такие протекающие промоторы присутствуют и в случае геометрии линейной последовательности. Кроме того, имеется конкуренция другого сорта: если два промотора перекрываются или очень близко расположены, то экспериментально установлено, что полимеразы, пытающиеся связаться с ними, мешают друг другу за счет диффузии в трехмерной окрестности этих промоторов. Здесь много и конкретных вопросов. Например, каковы средняя длина пройденного полимеразой участка и асимптотическое распределение этих длин.

3. Согласование набора деревьев (результат согласования — дерево видов). Хотя каждый ген вместе с его регуляторной системой развивается внутри вида, эволюция гена, системы и эволюция вида, как правило, далеки друг от друга. Фундаментальная задача состоит в переходе к непрерывному времени и к среде из генов, систем и видов. Мы рассмотрим более обычный подход: гены вместе с их системами эволюционируют в дискретном времени, как бы независимо друг от друга, а потом их нужно согласовать между собой относительно эволюции вида. Эволюция каждого элемента (гена, системы, гена-системы, вида) описывается своим деревом. Пусть эволюция гена задается деревом G_i («деревом гена»). Дан набор генов и соответствующих деревьев $\{G_i\}$. Найти дерево S («дерево вида»), которое в *среднем наиболее близко* к набору $\{G_i\}$. Программа решения этой задачи такова: каждому G_i сопоставить степень $c(G_i,S)$ отличия G_i от неизвестного S, а затем минимизировать функционал $c(\{G_i\},S)=\sum_i c(G_i,S)$ по переменной S. Определить $c(G_i,S)$ как число отличий в эволюционном развитии гена от эволюционного развития вида. Для этого нужно определить список эволюционных событий и сопоставить дискретное время, текущее по дереву G_i, с дискретным временем, текущим по дереву S. Последнее требует определить отображение вершин из G_i в вершины и ребра из S (получается «сценарий эволюции» гена G_i вдоль дерева видов S). Нами предложены решения этих задач, причем алгоритмами не более, чем кубической (т.е. очень низкой) сложности, которые доступны по адресу http://lab6.iitp.ru/ru/super3gl/. В них неизвестное дерево видов S вместе со сценариями эволюции генов строится индуктивно по мере возрастания мощности множества V листьев в S. А именно, на каждом шаге уже известны деревья S_1 (с множеством V_1 листьев) и S_2 (с множеством V_2 листьев) и соответствующие им наборы сценариев f_1 и f_2.

Эти деревья склеиваются в одно большее дерево S_1+S_2 с объединенными сценариями f_1+f_2 так, чтобы степень $c(\{G_i\},S_1+S_2)$ была минимальной относительно *всевозможных разбиений* V на две части V_1 и V_2. На той же идее основано построение сценария эволюции гена вдоль известного дерева S: роль меньших деревьев играют два поддерева в S. Эти поддеревья должны иметь корни, находящиеся в одном *временном слое*. Мы предложили алгоритм, который разбивает множество ребер в S на временные слои, так что между ребрами из одного слоя возможны одномоментные события. Однако остается проблема обоснования такого разбиения.

4. Реконструкция вторичной структуры вдоль дерева (на примере реконструкции аттенюаторной регуляции). Нам нужны представления о первичных и вторичных структурах, об аттенюаторной регуляции. Некоторые регуляторные участки (*первичные структуры*), будучи скопированы (т.е. оторваны от целого генома) образуют еще и *вторичные структуры* (ВС, рис. 1–3); каждая ВС состоит в спаривании букв A с T и G с C (водородной связью пар и специальной связью соседних пар). На рис. 1 показан один элемент («спираль») такой структуры. Биологические ВС содержат в той или иной комбинации до тысячи и более спиралей. Такое спаривание происходит участками («плечами») некоторой длины (на рис. 1 длины плеч 6, 3 и 4). Спираль состоит из нескольких «гипоспиралей» — связных спаренных участков: два *максимально продолженных без разрывов* плеча, соединенные своей петлей (на рис. 1 показано три вложенных друг в друга гипоспирали, каждая имеет свою петлю с длиной 25, 18 и 6). Перед определенными генами важны вторичные структуры только определенного типа. Один из типов называется *аттенюаторной регуляцией*, ее существенная часть (пара альтернативных спиралей) показана на рис. 2. Итак, дано дерево S (видов или белковых регуляторов) и каждому его листу приписана первичная структура. В ряде случаев из экспериментальных данных известны вторичные структуры, образующиеся в этих первичных структурах. Однако эти вторичные структуры не даны в задаче и далеко не всегда известны, их нахождение — цель задачи. Когда они известны, то используются для независимого контроля решения. Нужно найти *соответствующее эволюции* распределение (конфигурацию) структур: первичных во внутренних вершинах дерева S и вторичных во всех его вершинах. Наше решение основано на гиббсовском подходе с функционалом энергии $H(\sigma)$, глобальные минимумы которого должны описывать варианты искомой конфигурации σ'. Точки σ' глобального минимума находятся методом аннилинга на основе стохастической динамики Метрополиса-Хастингса. Сам функционал $H(\sigma)$ является суммой трех слагаемых. Первое слагаемое отражает энергию парного взаимодействия двух первичных структур на концах каждого из ребер в дереве S. Точнее, оно отражает стандартную динамику первичной структуры: вероятности замен букв согласно фиксированной матрице замен и вероятности вставок/стираний какого-то слова произвольной длины в произвольной позиции первичной структуры. Для каждой позиции скорость эволюции в ней определяется на основе гамма-распределения. Второе слагаемое отражает консервативность вторичной структуры вдоль каждого ребра и даже вдоль целых путей в дереве S с помощью сложного потенциала нелокального взаимодействия. Третье слагаемое отражает присутствие других элементов рассматриваемой регуляции (например, гена «лидерного пептида»). Первое и второе слагаемые требуют парного выравнивания: соответственно первичных и вторичных структур на концах ребра. Для этого мы развили процедуру выравнивания вторичных структур у двух первичных структур. Алгоритм аннилинга реализуется как неоднородная марковская цепь, переходные вероятности которой зависят от текущей конфигурации $\sigma(n)$ и параметра β_n, характеризующего условную температуру системы. Пусть последовательность конфигураций $\sigma(n)$ начинается с любой $\sigma(0)$ и $\beta_n \to \infty$ так, что $\lim(\log n/\beta_n) > C$. Тогда доказано, что $\sigma(n)$ сходится по вероятности к одной из минимальных конфигураций σ'; так описывается все их множество.

5. Конкуренция двух процессов (транскрипции и трансляции — аттенюаторная регуляция). Еще об аттенюаторной регуляции. По последовательности друг за другом движутся две молекулярные машины, одна — полимераза, другая называется рибосомой. Рибосома связывается со своим сайтом (аналогом промотора) перед специальным геном (геном «лидерного пептида») после того, как полимераза уже связалась со своим промотором и ушла вперед на некоторое расстояние. Если рибосома догоняет полимеразу, то рибосома снижает скорость и движется вслед за полимеразой, не влияя на нее. Скорость рибосомы по определенному закону $\nu(c)$ зависит от концентрации c некоторого вещества (аминокислоты), не превосходя 45 букв/сек. На участке последовательности между полимеразой и рибосомой формируется вторичная структура (ВС) ω с наименьшей энергией среди всех возможных, которая по определенному закону снижает скорость $\nu(\omega)$ полимеразы. При отсутствии ВС ее скорость 42 букв/сек. Если в какой-то момент пониженная скорость полимеразы сочетается с ее нахождением на участке, имеющем много букв T (тогда связь полимеразы с последовательностью слабеет), то эта связь разрывается, и полимераза покидает последовательность («терминация транскрипции»). Дана последовательность, по которой таким образом движутся полимераза и за ней рибосома. Найти зависимость $p(c)$ частоты терминации транскрипции от величины c. Обычно $\nu(c)$ опре-

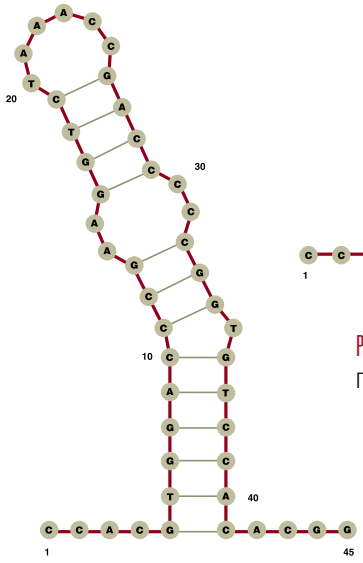

РИС. 1.

Спираль вторичной структуры, один из сотен составляющих структуру

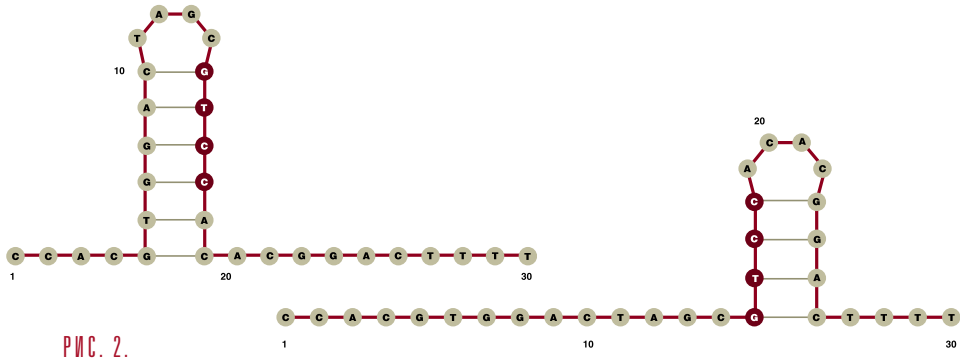

РИС. 2.

Пара альтернативных спиралей – существенная часть механизма регуляции работы гена

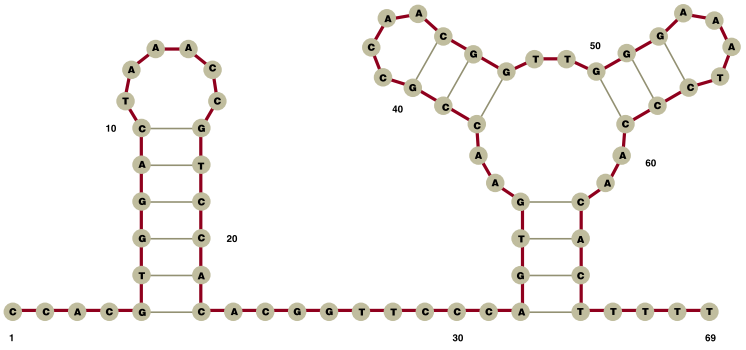

РИС. 3.

Пример записи вторичной структуры с помощью скобок

деляется по закону Микаэлиса-Ментен, вопрос о выборе $\nu(\omega)$ гораздо более сложный. Предложенное нами компьютерное решение доступно по адресу. Эта задача включает два вопроса, имеющих большое самостоятельное значение. По первому из них мало что известно: как определить силу сцепления молекулярной машины (полимеразы, рибосомы и т.п.) с последовательностью, по которой она движется? Каково влияние ВС на силу сцепления? Замечено, что эта сила убывает с уменьшением скорости движения полимеразы. Тогда: как ВС уменьшает скорость движения и как уменьшение скорости уменьшает силу? Напротив, по второму вопросу имеется много эмпирических исследований, но отсутствует теория. Как классифицировать ВС, биологически наблюдаются очень сложные ВС с множеством псевдоузлов; как приписать энергию данной ВС. Мало что известно о классификации псевдоузлов и о декомпозиции ВС на какие-то элементарные ВС. Рассмотрим простейший случай, когда ВС состоит из одной спирали (рис. 1). Напомним: спираль состоит из нескольких гипоспиралей. Мы приписывали спирали энергию по формулам: для энергии связи равную $\frac{1}{RT}\cdot\sum_i E_i$ и для энергии петель равную $\sum_i\left(1.77\cdot\ln(l_i+1)+B+\frac{C}{l_i}\right)$, где i пробегает все гипоспирали у спирали и E_i — энергия i-й гипоспирали, вычисляемая по таблицам водородной связи и связи соседних пар (стекинга); l_i – длина петли у i-й гипоспирали, а B и C — некоторые константы. Следующая трудная проблема: пространство всех ВС слишком велико, желательно разбить его на кластеры («макросостояния») и уже кластеру приписывать энергию. Это разбиение должно быть эффективным и в этой связи поступают следующим образом. Диаграмма — это скобочная структура, в которой каждая пара скобок соответствует гипоспирали и помечена номером соответствующей спирали, рис. 3. Скобочная структура понимается так: последовательная пара пар скобок $(\,)_1(\,)_2$... соответствует последовательно расположенным гипоспиралям; расположение первой гипоспирали в петле второй гипоспирали представляется вложенной парой пар скобок $((\ldots)_1)_2$. Так могут быть описаны и простые псевдоузлы: $(_1(_2)_1)_2$. Макросостояние — это множество всех ВС (которые соответствуют «микросостояниям»), соответствующих данной диаграмме; это множество предполагается непустым.

6. Сочетание 3-мерной и 1-мерной диффузий. Промотор имеет небольшую длину (до нескольких десятков букв), а типичная последовательность имеет несколько миллионов букв (у бактерии). Полимераза плавает в клетке, и перед началом ее движения по последовательности должна связаться со своим промотором (сильное, «специфическое» связывание). Как полимераза находит свой промотор? Последовательность (ДНК) расположена в клетке специальным образом, как кривая Жордана в соответствующем квадрате (у бактерий и архей). И ее геометрия играет важную роль. Существует следующее представление: специфическое связывание начинается с того, что полимераза связывается с ближайшим к ней участком последовательности слабой («неспецифической») связью и движется в одном из двух направлений

(случайно выбираемых) некоторое случайное короткое время. Это — одномерная диффузия полимеразы вдоль кривой. Затем полимераза отрывается (из-за слабой связи или столкновения) от последовательности и снова неспецифически связывается с другим участком кривой, который, если бы продолжать двигаться по кривой, расположен очень далеко от первого участка. Итак, после одномерной диффузии короткое время происходила трехмерная диффузия, а затем опять началась одномерная и т.д. до тех пор, пока полимераза не приблизится к своему промотору. Задача состоит в исследовании такого сочетания двух диффузий с учетом вида или только характеристик кривой. Здесь много экспериментальных данных, но теория, насколько нам известно, ограничена. Взаимодействие спиралей РНК с рибосомой, еще не начавшей трансляцию, но движущейся вдоль РНК в поисках инициирующего кодона, также связано с диффузией. Хотя сейчас реализовать моделирование диффузии слишком трудно, это позволило бы точнее определять инициирующие кодоны, в том числе, отличающиеся от обычного ATG.

7. Происхождение видов. Рассматривается характеристика генома, определяемая числовой последовательностью x, в которой на i-м месте находится число m_i разных генов, каждый из которых имеет ровно i копий (копия гена — также ген). Числа m_i — неотрицательные и все целые или все вещественные, а с некоторого места в x идут одни нули. Обозначим $m(x)=m_1+m_2+\ldots$ — *число всех типов* генов и $n(x)=m_1+2m_2+\ldots$ — *число всех генов* в геноме с характеристикой x. Пусть V — пространство всех допустимых последовательностей x и $f(x,t)$ — плотность геномов в точке x в момент времени t. Заметим, что в этой модели геномы и гены представлены только через их характеристики. Для точки x разрешены следующие переходы (соответствующие события происходят с генами и геномами).

1) $<\ldots,m_i,\ldots>\to<\ldots m_{i-1}+1, m_i-1,\ldots>$ *потеря* одного гена среди m_i, если $i\neq 1$ и $m_i\geq 1$, и $<\ldots,m_i,\ldots>\to<\ldots m_1-1, m_2,\ldots>$, если $i=1$ и $m_1\geq 1$; если $m_i=0$ или $m_1=0$, то этот переход запрещен. 2) $<\ldots,m_i,\ldots>\to<\ldots m_1+1, m_2,\ldots>$ *перенос*, т.е. появление нового гена, представленного одной копией. 3) $<\ldots,m_i,\ldots>\to<\ldots m_i-1, m_{i+1}+1,\ldots>$, $i\neq 1$ *дупликация* гена среди m_i; при этом $m_i\geq 1$, иначе переход запрещен. 4) $<\ldots,m_i,\ldots>\to<m_1+1,\ldots, m_{i-1}+1, m_i-1,\ldots>$ *мутация* гена среди m_i, если $i\neq 1$, и $<\ldots,m_i,\ldots>\to<\ldots,m_i,\ldots>$, если $i=1$; при этом $m_i\geq 1$, иначе переход запрещен. Для каждого из переходов определен свой вектор скорости (интенсивности) перехода, зависящий от точки x. Их сумму обозначим $A(x)$, она задает векторный потенциал. Скалярный потенциал определим как $-V(x)$, где $V(x)$ отражает внутреннюю согласованность («выживаемость») генома в точке x. Оба потенциала зависят от параметров, среди которых выделяются $m(x)$ и $n(x)$; некоторые параметры неизвестны и их предполагается варьировать. Пусть $V(x)$ принадлежит классу V функций, которые отличаются невысокими хаотично расположенными максимумами. Такие V соответствуют представлению: природа заранее не сделала выбора, какие геномы будут жизнеспособными в процессе их эволюции под действием векторного потенциала $A(x)$, но все-таки заложила в $V(x)$ небольшие предпочтения. Скалярный потенциал $V(x,t)$, вообще говоря, зависит еще от времени, т.е. сам подвержен некоторой динамике в пространстве V. Например, можно поставить вопрос так: в моменты времени t_i, определенные по пуассоновскому распределению с параметром μ, происходят катаклизмы. Это — достаточно резкие смены выживаемости, когда происходит переход от $V(t_i)$ к $V(t_{i+1})$, состоящий в перемещении и небольшом изменении локальных максимумов в $V(t_i)$ согласно некоторому распределению с одним параметром λ. Существует ли естественное распределение и значения параметров μ и λ, при которых с некоторого момента времени в пространстве V начинают формироваться кластеры (биологически — виды). Поясним последнее. Мы хотим описать область параметров, для которых существует момент времени t_0, начиная с которого траектории обладают свойством: «почти вся масса $M(t)=\int f(x,t)dx$ сосредотачивается в нескольких дизъюнктных кластерах», (*). Эти кластеры представляют характеристики возникших видов. Число кластеров можно заранее оценить через число известных видов, что послужит условием в задаче. Тогда t_0 представляет момент происхождения видов. Из численного моделирования известны значения параметров, при которых имеет место свойство (*). Мы не обсуждаем биологически более адекватную картину, в которой геном представлен более явным образом, как линейная последовательность натуральных чисел с повторениями, в которой каждое число — имя гена. В этом случае динамика генома получает более сложное описание.

7.1. Динамику характеристики $x=x(t)$ можно описать и по другому. А именно, уравнением $x'=A(x)+\varepsilon\xi$, где ξ — шум с некоторым генератором, определяемым потенциалами, и ε — параметр. Можно предположить, что существует момент времени t_0, начиная с которого имеется конечное число массивных кластеров с центрами масс $x_1,x_2,\ldots$, переходы между которыми требуют экспоненциально долгого времени или невозможны, (**). Тогда эти x_j — характеристики возникших видов, а t_0 — момент происхождения видов. В духе теории Вентцель-Фрейдлина можно найти функцию $\varphi(x)$, для которой равенство $\varphi'(x)=0$ является необходимым условием для выполнения (**). Тогда x можно находить, решая это уравнение.

Любецкий Василий Александрович
д.ф.-м.н, профессор, зав. лабораторией математических методов и моделей в биоинформатике Института проблем передачи информации им. А.А. Харкевича РАН

127994, г. Москва, Большой Каретный пер., д. 19,
тел.: +7 (910) 464-69-17, +7 (495) 413-46-43,
e-mail: http://lab6.iitp.ru/, lyubetsk@iitp.ru