В.А. Любецкий, А.В. Селиверстов

Математические задачи биологической эволюции и регуляции молекулярного уровня

Институт проблем передача информации РАН, Москва

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 задачи, что было бы очень желательно для 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 1-4 на уровне моделирования исследованы нами. Для них нами были построены численные компьютерные модели, которые дали результаты близкие к экспериментальным данным, обычно даже находящиеся в пределах экспериментальных ошибок. Напротив, для задач 5-6 не ясно даже, как эффективно проводить моделирование. Не всегда доступны прямые экспериментальные измерения величин, о которых говорится в задачах 1-6. Иногда модель приходится сравнивать с косвенными биологическими данными, что однако является обычной ситуацией. Биологические термины, упоминаемые ниже, не существенны для математического понимания задач 1-6. Следующий абзац носит *общекультурный характер*.
Геном – последовательность (длиной миллионы или миллиарды) букв в 4х буквенном алфавите $\{A, T, G, C\}$, а ген – направленный участок в этой последовательности, т.е. также последовательность букв, но гораздо более короткая. Число генов может быть до 8-9 тысяч (у бактерий) и до 50 тысяч (у животных). Число генов может быть и малым: несколько сот (у бактерий, пластид) или несколько штук (у митохондрий). Между генами расположены участки, которые регулируют активность (интенсивность работы) генов. Как и всё в геноме, они являются направленными последовательностями (более короткими, чем гены) букв и кодируют сложные структуры. Таким образом, геном – это множество генов и регуляторных участков («регуляторных систем»). 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. Организм – это геном (в его прижизненном развитии), а вид – это совокупность геномов с близкими характеристиками, что позволяет виду иметь потомство, которое также состоит из геномов. Все процессы развиваются в физическом времени, хотя до сих пор в моделях принято для упрощения (возможно, кажущегося) рассматривать дискретное время.

1. Конкурирующие процессы связывания и движения (конкуренция РНК-полимераз). Дана последовательность в 4х буквенном 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. 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 (например, экспоненциально распределенное) и в мгновенном возвращении в исходное положение. Такие односторонние колебания продолжаются случайное число раз (например, с геометрическим распределением) 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, а 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 (комбинации аффинных функций) 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 http://lab6.iitp.ru/ru/rivals/. 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 рассматривать окружность, т.е. последовательность по модулю ее длины. Самый простой пример – это конкуренция на окружности 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 будет 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 ещё «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.

2. Согласование между собой набора деревьев (результат согласования – дерево видов). Гены и регуляторные системы – части организма (вида); соответственно организм (вид) – совокупность генов и их регуляторных систем. 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 здесь к непрерывному времени и к среде из генов, систем и видов. Рассмотрим более обычный подход: гены вместе с их системами эволюционируют в дискретном времени, как бы независимо друг от друга, а потом согласуются 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. Переход к непрерывному времени может снять ее.

3. Реконструкция вторичной структуры вдоль дерева (на примере реконструкции аттенюаторной регуляции). Некоторые регуляторные участки (*первичные структуры*), будучи скопированы (оторваны от целого генома) образуют еще и *вторичные структуры* (ВС, см. рис. 1-3); каждая ВС состоит в спаривании букв A с T и G с C (водородной связью пар и стекингом 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. Дано дерево *S* (видов или факторов регуляции) 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. Итак, 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. The functional $H(\sigma)$ is a sum of three terms. The first term reflects the energy of pair interaction between the two next primary structures on each edge. More exactly, it reflects the standard dynamics of the primary structure: the probability of a letter substitution according to the fixed transition rate matrix, and also the probabilities of insertion/deletion of a word with any length at any position in the primary structure. At each position, the evolution rate is considered according to the gamma law. The second term reflects the conservativity of secondary structure along each edge and entire paths in the tree *S* that is specified by a sophisticated potential of non-local interaction. The third term reflects the presence of other elements pertinent to the regulation of interest (e.g., the "leader peptide gene"). The first and second terms require a pairwise alignment to be found: primary structures at the ends of each edge to be aligned for the first term computation, and secondary structures – for the second term. For the latter, we have developed a procedure that aligned the secondary structures of two primary structures. Алгоритм реализуется как неоднородная марковская цепь, переходные вероятности которой зависят от текущей конфигурации $\sigma(n)$ и от параметра β_n, характеризующего температуру. Пусть последовательность конфигураций $\sigma(n)$ начинается с любой $\sigma(0)$ и $\beta_n \to \infty$ так, что $\lim(\log n / \beta_n) > C$. Тогда $\sigma(n)$ сходится по вероятности к одной из минимальных конфигураций σ', и таким образом описывается всё их множество, http://lab6.iitp.ru/ru/anneal/.

4. Конкуренция двух процессов (транскрипции и трансляции – аттенюаторная регуляция). По последовательности друг за другом движутся две молекулярные машины, одна – полимераза, другая называется рибосомой. Рибосома связывается со своим сайтом (аналогом промотора) перед специальным геном («ген лидерного пептида») после того, как полимераза уже связалась со своим промотором и ушла вперед на некоторое расстояние. Если рибосома догоняет полимеразу, то рибосома снижает (если она была выше) скорость и движется вслед за полимеразой, не влияя на нее. Скорость рибосомы по определенному закону $v(c)$ 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 $v(\omega)$ снижает скорость 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), тогда 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 $v(c)$ выбирается according to the Michaelis-Menten law, вопрос о выборе

$v(\omega)$ гораздо сложнее. Предложенное нами компьютерное решение доступно по адресу http://lab6.iitp.ru/rnamodel/runmodel.php?lang=rus. Эта задача включает два вопроса, имеющих большое самостоятельное значение.

4.1. 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.

4.2. Напротив, по второму вопросу имеется много эмпирических исследований, но отсутствует теория. Как классифицировать ВС, биологически наблюдаются очень сложные ВС с множеством псевоузлов; 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. Тогда мы приписывали спирали энергию по формулам: bond energy as follows: $\frac{1}{RT}\cdot\sum_i E_i$ and loop energy as follows $\sum_i\left(1.77\cdot\ln(l_i+1)+B+\frac{C}{l_i}\right)$, , where i varies over all hypohelices of the helix и E_i – энергия i-th hypohelices, вычисляемая по таблицам водородной связи и стекинга; l_i – длина петли у i-th hypohelices, 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. A diagram is an parentheses structure with each pair of parentheses corresponding to a hypohelix and tagged with the number of its parental helix, fig. 3. The parentheses structure is understandable as follows: consecutive hypohelices correspond to consecutive pairs of parentheses, $(\,)_1(\,)_2...$; overlapping of first hypohelix with the loop of second hypohelix is represented by embedded structure $((...)_1)_2$. 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.

5. **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. И ее геометрия играет важную роль. Существует 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 двигаться по кривой, был расположен очень далеко от первого участка. Итак, 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 известно, ограничена.

6. 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 $<\ldots, m_i, \ldots> \to <m_1 - 1, m_2, \ldots>$, если $i = 1$ и $m_1 \geq 1$; если $m_i = 0$ или $m_1 = 0$, то этот переход запрещен. 2) $<\ldots, m_i, \ldots> \to <m_1 + 1, m_2, \ldots>$ *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 $\boldsymbol{V}$ 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 $\boldsymbol{V}$. Вопрос о точном выборе этого класса потенциалов не вполне ясен, как и вопрос о выборе этой динамики. 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 распределения 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.
6.1. 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_j можно находить, решая это уравнение.

Авторы благодарны Л. Русину, К. Горбунову, Л. Рубанову, С. Пирогову, Е. Жижиной за подробное обсуждение всего текста и за участие в качестве соавторов в решении некоторых из перечисленных задач. Авторы также благодарны Л. Русину за помощь в подготовке этого текста.

Ниже в естественном порядке приведены рисунки 1-3.