Информационные процессы, Том 5, № 5, 2005, стр. 337–366.

Алгоритм и результаты счета для модели регуляции экспрессии генов у бактерий на основе формирования вторичных структур РНК[1]

В.А.Любецкий, К.Ю.Горбунов, С.А.Пирогов, Л.И.Рубанов, А.В.Селиверстов

Институт проблем передачи информации РАН, Москва, Россия
lyubetsk@iitp.ru, rubanov@iitp.ru

Поступила в редколлегию 16.09.2005

Аннотация—В работе[2] предлагается модель, в первую очередь, классической РНКовой регуляции экспрессии генов с помощью прерывания (терминации) процесса транскрипции. Модель опирается на представление о макросостоянии вторичной структуры в регуляторной области РНК между рибосомой и полимеразой, на формулы резонансного типа, определяющие величину замедления РНК-полимеразы набором шпилек в той же области, на представления о процессах посадки и последующего движения рибосомы и полимеразы. Специальное внимание уделяется подбору параметров модели. Для проверки модели проведено компьютерное моделирование и получены, в частности, зависимости вероятности терминации транскрипции от величины концентрации загруженных тРНК и от концентрации аминокислоты в клетке или в культуре для многих регуляторных областей в геномах бактерий (здесь данные приводятся для четырех стрептомицетов) и при различных значениях трех параметров, которые авторы рассматривают как основные. Полученные зависимости согласуются с доступными экспериментальными данными; в том числе, по форме графиков, относящихся к активности фермента в зависимости от концентрации аминокислоты (например, атранилат синтазы от триптофана в культуре у *S. venezuela*). Белок-ДНКовая регуляция транскрипции, как и секвестор трансляции, находят отражение в предлагаемой модели, но их подробные разработки будут представлены в другой статье. В дальнейшем на основе нашей модели предполагается получить предсказания о влиянии точечных “мутаций” в регуляторных областях геномов на результат аттенюаторной регуляции, включая предсказания об эволюционной устойчивости организмов. А затем — включить эту модель в более широкую модель регуляции и метаболизма у бактерий. Другое возможное использование: сейчас аттенюаторная регуляция предсказывается обычно на основе множественного выравнивания, для этого требуется несколько последовательностей; получение с помощью модели на индивидуальной последовательности характерной для аттенюации или ее отсутствия кривой при подходящих параметрах могло бы рассматриваться как аргумент в пользу наличия или отсутствия аттенюации.

ВВЕДЕНИЕ

Такая модель зависит от многих достаточно произвольных решений: о декомпозиции всего процесса на составные части, о выборе математического аппарата для описания частей, о выборе списка и численных значений параметров, о способе сопоставления результатов моделирования с пока малочисленными экспериментальными данными, и т.д. В частности, не представляется разумным, по крайней мере, сразу описывать с одинаковой подробностью составные части этого процесса; поэтому приходится выбирать еще и степень подробности в

[1] Работа выполнена при частичной поддержке гранта МНТЦ № 2766.
[2] Биологическая часть этой публикации представлена в журнал “Молекулярная биология”.

описании каждой из них. Авторы видят путь преодоления этих трудностей в обсуждении и сравнении предлагаемых моделей между собой и с экспериментальными данными. Можно надеяться, что обсуждение таких моделей могло бы стимулировать соответствующие эксперименты.

Модель реализована в виде 32-разрядной компьютерной программы на языке C++ в стандарте ANSI и имеет интерфейс командной строки с передачей исходных данных и результатов через текстовые ASCII файлы. Это позволяет проводить разнообразные компьютерные эксперименты на различных платформах с любыми исходными последовательностями и с большим числом доступных пользователю как вычислительных, так и биологических параметров процесса.

В работах [1]-[4] А.А. Миронова и коллег рассмотрено моделирование по Монте-Карло процессов РНКовой терминации на уровне микросостояний, а также была поставлена задача моделирования этого процесса на уровне макросостояний. Кроме определения макросостояния (где мы, в основном, следуем [1]-[4]), наша задача состояла в выводе формулы для замедления РНК-полимеразы набором шпилек и, в конечном счете, текущим макросостоянием на участке РНК между рибосомой и полимеразой, а также — в представлении инициации и элонгации процессов транскрипции и трансляции.

Приведенное ниже определение макросостояния можно сужать путем ограничений на его естественные параметры (такие, как числа спаренных и неспаренных нуклеотидов, длины петель и выпячиваний, и т.д.) или, наоборот, расширять, объединяя в гиперсостояние те из них, между которыми возникает большое число быстрых взаимных переходов. Аналогично приведенная ниже "резонансная" формула для взаимодействия шпильки и полимеразы может быть выписана более детально и на основе более сложного (даже нелинейного) уравнения — пока это сделано на минимальном уровне, допускающем сравнение с экспериментом. Влияние таких процессов как репрессия и активация, срыв полимеразы за счет взаимодействия со специальными белками пока представлено в простейшей форме, ступенчатой функцией, просто увеличивающей долю случаев, в которых происходит терминация. Величина, управляющая всем этим процессом, — концентрация аминокислоты рассматривается в конечном счете в культуре, так как только для этого случая имеются экспериментальные зависимости активности фермента от концентрации аминокислоты.

Непосредственно в статье рассматривается классическая аттенюаторная регуляция, но аналогичным образом нами рассмотрены и другие РНКовые регуляции, например, получено среднее время перекрывания области Шайн-Дельгарно шпильками, которые препятствуют посадке на нее рибосомы, рассмотрена регуляция с помощью вторичных структур, включающих Т-бокс, рибосвич и т.д.

1. ОПИСАНИЕ МОДЕЛИ

1а. Определения микро- и макросостояний. Вывод формулы для константы скорости переходов между ними.

Предполагается, что дана и везде далее фиксирована последовательность в четырехбуквенном алфавите $\{a, c, t, g\}$ — регуляторная область в геноме бактерии или случайная последовательность. Например, в одном из вариантов биологическая область вырезалась от старта лидерного белка до конца поля остатков урацила с характерными длинами: лидерный пептид до 100 нуклеотидов (например, 99 у *ilvL* в *E. coli*, 45 у *trpL* в *E. coli*, 57 у *trpE* в *Streptomyces*, 30 нуклеотидов у гипотетического лидерного пептида перед геном *cbs* у *B. longum*, везде включая стоп-кодон), поле урацилов до 8 букв T, длина всей области до 200 нуклеотидов. В том же варианте "начальное" состояние моделирования можно представить себе так, что рибосома находится ее активным центром на старт кодоне (atg или gtg) лидерного пептида.

В исходной последовательности выделяются отрезки длиной не менее 3 нуклеотидов — плечи будущих спиралей: $\ldots a_i, \ldots, b_j, \ldots$. При спаривании каких-то отрезков a_i и b_j одинаковой длины (подразумевается образование водородных связей между соответствующими нуклеотидами вдоль всей длины двух отрезков a_i и b_j) получается спираль γ_i — везде предполагается, что спираль γ_i *непродолжаемая* и промежуток между отрезками (петля спирали) имеет длину не менее 3 нуклеотидов. Комплементарными, т.е. такими, для которых возможно спаривание, считаются пары нуклеотидов: gc и at с сильной связью, gt со слабой связью; пары нуклеотидов ac, ag, ct не считаются комплементарными. Список пар нуклеотидов, которым разрешено образовывать между собой водородную и ван-дер-ваальсову связь (синоним: спариваться), является параметром алгоритма. Ван-дер-ваальсова связь образуется между соседними парами спаренных нуклеотидов (стекинг); именно она вносит основной вклад в вычисление энергии. Алгоритмы допускают, вообще говоря, любой список *исходных спиралей*; выше определен лишь один из возможных вариантов, в котором в качестве исходных берутся все непродолжаемые спирали с указанными ограничениями на плечо и петлю.

Все эти представления, как и описание самой классической аттенюаторной РНКовой регуляции экспрессии генов в зависимости от величины концентрации определенного вещества (аминокислоты или загруженной тРНК; последняя, в свою очередь, определяется концентрациями аминокислоты и аминоацил-тРНК синтетазы), изложены, например, в [5, с. 172-189].

Гипоспиралью спирали γ_i называется любая непустая часть $\bar{\gamma}_i$ спирали γ_i, состоящая из двух связных плеч длины *не менее* 3 нуклеотидов. Здесь и далее *плечами* называются спариваемые отрезки гипоспирали или спирали, концы которых будут стандартно *обозначаться* (считая от $5'$-начала исходной последовательности) буквами A, B, C, D. Петлей называется участок цепи РНК между двумя плечами гипоспирали.

Микросостоянием называется (непустой совместный) набор гипоспиралей, непродолжаемых *в этом наборе* и без псевдоузлов, для которого никакие две гипоспирали не соприкасаются (т.е. A и D одной из них не являются оба соседними нуклеотидами к B и C другой из них); кроме того, отдельным "начальным" микросостоянием является пустое множество $\emptyset$. *Псевдоузлом* называется пара гипоспиралей, у которой ровно одно плечо одной из них пересекается с петлей другой (и, следовательно, находится в этой петле). Все гипоспирали от одной спирали, вошедшие в данное микросостояние, называются *подспиралью* этой спирали в данном микросостоянии. Подспираль (данного микросостояния) однозначно разлагается на гипоспирали (этого микросостояния).

Для любого микросостояния каждая из его гипоспиралей и подспиралей получает тот же номер, что и (непродолжаемая) спираль, из которой она взята; при этом все спирали исходной последовательности нумеруются в каком-то одном заранее фиксированном порядке.

Диаграммой микросостояния называется обычная скобочная структура, отражающая взаиморасположение всех гипоспиралей в микросостоянии, а каждой паре скобок (по другой терминологии: *хорде*) приписывается номер той спирали, из которой взята гипоспираль, соответствующая паре скобок (хорде). Опуская в диаграмме левую скобку, мы, естественно, опускаем соответствующую ей правую скобку и соответствующий номер спирали, и снова получаем диаграмму. Скобочная структура отражает взаиморасположение гипоспиралей в соответствии с обычным правилом: нескольким расположенным друг за другом гипоспиралям соответствует такое же число последовательных скобок $(\cdot)(\cdot)\ldots(\cdot)$, а гипоспирали 1 в петле гипоспирали 2 соответствует вложение скобок $((\cdot))$, где внутренней паре скобок соответствует гипоспираль 1 и внешней паре скобок — гипоспираль 2. Иногда ту же скобочную структуру называют набором хорд, соединяющих плечи гипоспиралей этого микросостояния, а хордам приписаны номера соответствующих спиралей. Важно заметить, что в диаграмме номера спиралей могут неоднократно повторяться, так как из одной спирали могут быть взяты многие гипоспирали. По

микросостоянию, т.е. в сущности по списку всех спаренных нуклеотидов, легко выписывается его диаграмма. Но по диаграмме микросостояния нельзя восстановить само микросостояние: диаграмма сохраняет только "геометрию" взаиморасположения гипоспиралей и указание для каждой пары скобок, из какой спирали разрешено брать гипоспираль для этой пары скобок.

Носителем микросостояния называется набор спиралей, построенных по каждой гипоспирали этого микросостояния. Наоборот, любому *носителю* (т.е. произвольному набору непродолжаемых спиралей) соответствует множество микросостояний, "реализующих" его. А именно, каждому носителю, состоящему из спиралей $\gamma_1, \ldots, \gamma_k$, соответствует множество *реализующих его микросостояний*: это любой нерасширяемый (в себе) набор *подспиралей* $\bar{\gamma}1 \subseteq \gamma_1, \ldots, \bar{\gamma}_k \subseteq \gamma_k$ (причем от каждой спирали γ_i берется ровно один непустой и не обязательно связный участок $\bar{\gamma}_i$) без псевдоузлов. Как и выше, *соседние* гипоспирали (т.е. у которых пара A и D нуклеотидов расположена непосредственно вслед за парой B и C) объединяются. По произвольному носителю $\gamma_1, \ldots, \gamma_k$ можно эффективно проверять, соответствует ли ему непустое множество микросостояний. Это можно делать, например, с помощью алгоритма 1, представленного в приложении.

Макросостоянием называется любая *непустая* диаграмма; "непустая" в том смысле, что она имеет хотя бы одно реализующее ее микросостояние. Т.е. *макросостояние* — это скобочная структура (скобочная запись), в которой каждой паре соответствующих скобок приписано натуральное число — номер спирали в некоторой фиксированной нумерации всех спиралей исходной последовательности, имеющая хотя бы одну реализацию. Скобочная структура накладывает ограничение на взаимное расположение гипоспиралей, а номер спирали говорит, из какой спирали разрешено брать гипоспирали для этой пары скобок. Повторим: несколько скобок, расположенных подряд внутри одной скобки или вообще без объемлющей скобки, указывают на необходимость последовательного слева направо расположения соответствующих гипоспиралей, а расположение одной скобки внутри другой указывает, что первая гипоспираль находится внутри петли второй. Макросостоянию соответствует как носитель (т.е. набор всех к нему приписанных спиралей), так и список всех тех микросостояний этого носителя, которые имеют гипоспирали, расположенные в соответствии с его скобочной структурой, т.е. всех микросостояний, *реализующих* это макросостояние.

Таким образом, *макросостояние* — это такая диаграмма, что у любого из реализующих его микросостояний (хотя бы одно такое существует) диаграмма уже микросостояния совпадает с исходной диаграммой макросостояния.

По любой диаграмме эффективно проверяется, непустая ли она и, если так, то выписывается список всех микросостояний исходного макросостояния. Это делается с помощью алгоритма 2, указанного в приложении.

Энергия связи $E_{\bar{\gamma}_i}$ гипоспирали $\bar{\gamma}_i$ получается суммированием энергий связи всех последовательных пар ее спаренных нуклеотидов на основе стекинга — энергии связи соседних пар, может быть, еще с учетом особенностей концевых пар и т.п. Возможны разные формулы для вычисления такой энергии — они являются параметром алгоритма, и здесь не обсуждаются. По определению энергия связи — отрицательное число, зависящее только от самой гипоспирали.

Каждой гипоспирали $\bar{\gamma}_i$ из данного микросостояния ω приписывается число l_i нуклеотидов в ее петле, не вошедших в петли и плечи других гипоспиралей из этого микросостояния, расположенных в петле $\bar{\gamma}_i$. Это число называется *длиной петли* гипоспирали $\bar{\gamma}_i$. Оно зависит от всего микросостояния; возможны более изощренные формулы для подсчета "длины петли", здесь мы не обсуждаем этот вопрос.

Микросостоянию ω по определению приписываются две энергии: свободная *энергия петель* и свободная *энергия связи* гипоспиралей. Возможны разные варианты этих формул. Напри-

мер, такой: первая из них равна

$$G_{loop}(\omega) = \sum_i (A \cdot l_i + 1.77 \cdot \ln(l_i + 1) + B + \frac{C}{l_i} + \frac{d}{l_i^2} + \ldots), \tag{1}$$

где i пробегает все гипоспирали из ω, $G_{loop}(\omega) \geq 0$. Пока принимаем: $A = 0$, так как отрицательно заряженная нуклеотидная цепочка находится в среде с положительно заряженными ионами и, в результате, более или менее нейтрализуется (нет кулонового потенциала), $B = 5$ для петель и $B = 0$ для выпячиваний. Слагаемое $1.77 \cdot \ln$ — из центральной предельной теоремы с поправкой Флори, $\frac{C}{l}$, $C =$ 5-10 — как отражение того, что маленьким петлям трудно изгибаться (некая упругость, жесткость цепи мРНК), $d = 0$ и то же для последующих членов.

А вторая энергия равна

$$G_{hel}(\omega) = (\sum_j E_{\bar{\gamma}_j})/kT < 0, \tag{2}$$

при некоторых значениях энергий и температуры.

Эти формулы также являются параметром алгоритма и могут меняться. Здесь kT равно $1,38 \cdot 10^{-23}$ Дж/°К, умноженное, например, на 313°К, что соответствует примерно 40°С. При практическом счете $G_{hel}(\omega)$ умножается на число Авогадро (примерно $6 \cdot 10^{23}$), так что подсчитываются соответственно удельная энергия в числителе и универсальная газовая постоянная в знаменателе, т.е.

$$G_{hel}(\omega) = (\sum_j E_{\bar{\gamma}_j})/RT < 0, \tag{3}$$

где j пробегает все гипоспирали из ω.

В некоторых организмах представляется целесообразным учитывать кроме энергии связи гипоспирали еще энергию, ответственную за то, что для обеспечения распада гипоспирали нужно произвести частичное “раскручивание” петли гипоспирали в цитоплазме клетки, разорвать связи третичной структуры. С этой целью используется следующая поправка к формуле (3):

$$G_{hel}(\omega) = \frac{\sum\limits_j \left(E_{\bar{\gamma}_j} - \alpha \cdot \frac{l'_j}{\left(1 + \frac{l'_j}{l_{max}}\right)} \right)}{RT}, \tag{4}$$

где l'_j — *полная длина петли* гипоспирали $\bar{\gamma}_i$ (т.е. без учета спариваний в ее петле, принадлежащих другим гипоспиралям), $l_{max} = 10$, $\alpha =$ 0-10.

Переходы между микросостояниями делятся на “быстрые” и “медленные”. *Быстрый* переход — это по определению переход без изменения соответствующего макросостояния. *Медленный* переход — это по определению переход, при котором макросостояние меняется ровно на одну хорду, т.е. меняется на ± 1 хорду. При любом переходе разрешается перестройка, вообще говоря, любого числа гипоспиралей.

Возможны разные естественные варианты динамики как микросостояний, так и макросостояний, ниже приведем один из них, простейший.

Вероятность (далее везде вместо этого будем говорить — *скорость*) быстрого перехода из микросостояния ω в микросостояние ω' с *тем же макросостоянием* зададим формулой

$$K(\omega \to \omega') = \kappa_f \cdot \exp((G_{loop}(\omega) + G_{hel}(\omega)) - (G_{loop}(\omega') + G_{hel}(\omega'))),$$

если $(G_{loop}(\omega) + G_{hel}(\omega)) < G_{loop}(\omega') + G_{hel}(\omega')$, и

$$K(\omega \to \omega') = \kappa_f,$$

если $(G_{loop}(\omega)+G_{hel}(\omega)) \geq G_{loop}(\omega')+G_{hel}(\omega')$. Здесь вероятен переход к меньшей энергии. Константа κ_f называется "константой быстрого замыкания", она характеризует скорость быстрых переходов. Естественно ожидать, что выполняется неравенство $\kappa_f \geq \kappa$, где κ участвует в нижеуказанных формулах и называется "константой медленного замыкания", она характеризует скорость медленных переходов. Хотя все алгоритмы не зависят от этого неравенства.

Для медленных переходов между микросостояниями, когда *макросостояние обязательно меняется ровно на одну хорду*, примем следующие формулы. Скорость медленного перехода в случае **распада гипоспирали**, т.е. когда происходит уменьшение макросостояния на одну хорду, из микросостояния $\omega = \{\bar\gamma_{1i}, \ldots, \bar\gamma_{ki}\}$ (где указаны все его гипоспирали) в микросостояние $\omega' = \{\bar\gamma'_{1i}, \ldots, \bar\gamma'_{ki}\}$ (где также указаны все его гипоспирали, и $\bar\gamma'_{li} = \emptyset$ для какого-то одного i), гипоспирали $\bar\gamma_{li}, \bar\gamma'_{li}$ взяты от одной и той же спирали γ_l и соответствуют одной хорде (фактически она отсутствует в ω'), задается формулой

$$K(\omega \to \omega') = \kappa \cdot \exp(G_{hel}(\omega) - G_{hel}(\omega')). \tag{5}$$

Здесь вероятнее отпадение гипоспирали с маленьким плечом. Энергия связи уменьшается по абсолютной величине.

Скорость обратного перехода, т.е. медленного перехода в случае **присоединения гипоспирали**, т.е. когда происходит увеличение макросостояния на одну хорду, задается формулой

$$K(\omega' \to \omega) = \kappa \cdot \exp(G_{loop}(\omega') - G_{loop}(\omega)). \tag{6}$$

Здесь вероятно присоединение гипоспирали с большим плечом. Значение константы κ медленного замыкания, следуя статьям [1]-[4], принимается равным $\kappa = 10^6$-$10^7 \mathrm{c}^{-1}$.

Таким образом, если разрешить только быстрые переходы, то на множестве всех микросостояний ω при $t \to \infty$ установится следующее стационарное распределение вероятностей

$$p(\omega) = \frac{\exp(-(G_{loop}(\omega) + G_{hel}(\omega)))}{z(\Omega)}, \text{ где } z(\Omega) = \sum_{\omega \in \Omega} \exp(-G_{loop}(\omega) - G_{hel}(\omega)). \tag{7}$$

Если теперь описать динамику макросостояний на основе динамики реализующих их микросостояний, то возможны только два перехода: добавление к текущему макросостоянию Ω новой гипоспирали (хорды) γ и исчезновение из Ω одной из бывших в нем гипоспиралей (хорд) γ. После очевидного усреднения по всем парам микросостояний $\omega \in \Omega$, $\omega' \in \Omega'$ получим следующую формулу для *скорости перехода из одного макросостояния* Ω в другое макросостояние Ω':

$$K(\Omega \to \Omega') = \sum_{\omega \in \Omega} \sum_{\omega' \in \Omega'} p(\omega) \cdot K(\omega \to \omega'). \tag{8}$$

Эта формула относится как к случаю увеличения макросостояния, так и к случаю его уменьшения на одну гипоспираль. Авторами предложены эффективные реализации частей алгоритма, в частности, способ вычисления этих сумм, не предполагающий перебор всех пар микросостояний (см. алгоритм 3 в приложении).

Вопрос об обоснованности наших определений "быстрого" и "медленного" переходов находит некоторое подтверждение в следующем утверждении.

Предложение 1. Пусть даны два микросостояния, реализующих одно макросостояние (что эквивалентно *изоморфизму деревьев микросостояний*: ребрам деревьев приписаны гипоспирали, которые считаются эквивалентными, если берутся из одной спирали, а порядок непосредственных потомков каждой вершины фиксирован и сохраняется при изоморфизме). Тогда от одного микросостояния к другому можно перейти, оставаясь внутри макросостояния, цепочкой "шагов" так, что каждый шаг включает не более двух разрывов и двух рождений пар спаренных нуклеотидов.

Доказательство. Обозначим первое микросостояние v_1, а второе v_2. Будем представлять микросостояния в виде дерева, каждому ребру которого приписана гипоспираль (очевидно, что скобочной записи естественным образом соответствует нарисованное на плоскости и растущее, скажем, вверх дерево). Для каждого ребра r рассмотрим две гипоспирали — $v_1(r)$ и $v_2(r)$, приписанные этому ребру, соответственно, в микросостояниях v_1 и v_2.

Если $v_1(r)$ и $v_2(r)$ имеют в пересечении менее трёх склеек, *ориентируем* это ребро в направлении от $v_1(r)$ к $v_2(r)$ (то есть ориентируем ребро от корня, если в $v_2(r)$ есть склейки, лежащие в петле $v_1(r)$, и к корню, если наоборот), иначе оставляем ребро *неориентированным*.

Возьмем вершину ω, из которой нет исходящих ориентированных рёбер. Для неё совершаем переход от v_1 к v_2, укорачивая одни примыкающие к ω гипоспирали и удлиняя другие (с того конца, что примыкает к ω). Все пары склеенных в текущем состоянии нуклеотидов, которые либо есть в v_2, либо расположены "дальше" (от ω) тех, что в v_2, мы автоматически замораживаем и больше не расклеиваем.

Говоря нестрого, далее мы будем "причёсывать" ещё "непричёсанные" гипоспирали, т.е. такие, которые с *рассматриваемого* конца либо короче, чем надо, либо длиннее. Отметим также, что в процессе урезания гипоспираль не может укоротиться до длины меньше трёх, поскольку, в силу отсутствия исходящих рёбер как минимум три склейки уже заморожены. Заметим ещё, что как только очередная гипоспираль причёсана, то в дальнейшем процессе она останется непродолжаемой, поскольку даже если она теоретически продолжаема, то хоть одна из (двух возможных) примыкающих к ней гипоспиралей тоже становится причёсанной (и упирающейся в первую).

Поскольку все гипоспирали не могут быть короче, чем надо (иначе существовала бы продолжаемая спираль), то если есть хоть одна непричёсанная гипоспираль, то есть и такая гипоспираль g, что её крайней пары нет в v_2, т.е. g длиннее, чем надо (напомним, что у нас фиксирована вершина дерева и термины "короче", "длиннее", "крайняя пара" и т.д. мы релятивизуем относительно рассматриваемого конца). Обозначим эту пару через p. Найдём минимально возможное множество следствий расклейки p. Заметим, что g может упираться парой p не более, чем в две спирали. Обозначим их g_1 и g_2, а их пары, примыкающие к p — через p_1 и p_2. Возможны следующие случаи.

1. Каждая из гипоспиралей g_i либо в принципе непродолжаема, либо упирается парой p_i в ещё одну гипоспираль r_i (каждая в свою или в общую), отличную от g (т.е. обе g_i остаются непродолжаемыми после расклейки пары p). Выбираем такое i, что реализация гипоспирали g_i в v_2 длиннее, чем сейчас (оно есть, иначе в v_2 g была бы продолжаемой). Возможны следующие подслучаи:

1а) Гипоспираль r_i упирается ещё в некоторую гипоспираль t, которая после трёх событий (укорочение g, удлинение g_i и укорочение r_i) становится продолжаемой (заметим, что в силу выбора i крайняя пара спирали r_i не может быть замороженной). Тогда минимальное множество следствий состоит из 4 событий: указанных трёх и удлинения t.

1б) Гипоспираль t отсутствует или после упомянутых трёх событий остаётся непродолжаемой. Тогда минимальное множество следствий состоит лишь из трёх упомянутых событий.

2. Ровно одна гипоспираль g_i после расклейки p остаётся непродолжаемой. В качестве минимального множества следствий можно взять два события: укорочение g и удлинение g_j, где j не равно i.

3. Обе гипоспирали становятся продолжаемыми. Тогда событий три: укорочение g и удлинение g_1 и g_2.

Таким же образом продолжаем расклейку гипоспирали g, пока она не станет причёсанной.

Итак, во всех случаях можно переходами с не более чем двумя расклейками и не более чем двумя склейками перейти в более близкое к v_2 микросостояние (в смысле числа причёсанных гипоспиралей). После того, как все примыкающие к вершине ω концы гипоспиралей причёсаны, объявляем вершину ω обработанной и производим заново ориентировку рёбер, беря в качестве v_1 текущее микросостояние. Очевидно, новая ориентировка будет отличаться от старой лишь тем, что при этом не будет ориентированных рёбер, инцидентных ω (ни входящих, ни выходящих). Так что будет существовать необработанная вершина ω_1, из которой нет исходящих ориентированных рёбер. Обрабатываем её описанным образом и т.д. Когда все вершины будут обработаны, мы достигнем v_2. **Утверждение доказано**.

Заметим, что без предположения об изоморфизме деревьев, утверждение 1 неверно. Действительно, пусть имеются две спирали, из которых можно сделать выборку упирающихся друг в друга гипоспиралей (по одной из каждой) двумя способами, различающимися тем, какая гипоспираль находится в петле другой. Тогда мы можем (насколько позволяет длина) сдвигать границу между ними, но не можем без существенных усилий "поменять их местами". Например, пусть первая спираль состоит из 8 пар склеенных нуклеотидов, обозначаемых (от корня) как 1-1′, 2-2′,:, 8-8′, а вторая спираль — из пяти пар: 3-2′, 4-3′, 5-4′, 6-5′, 7-6′. Тогда есть два варианта: либо из первой спирали взять гипоспираль 1-1′, 2-2′, 3-3′, а из второй 5-4′, 6-5′, 7-6′, либо из первой взять гипоспираль 6-6′, 7-7′, 8-8′, а из второй 3-2′, 4-3′, 5-4′. Но переход от одного из этих микросостояний к другому требует, очевидно, двух шагов, каждый из которых состоит из пяти событий: один шаг — из трёх расклеек и двух склеек, а другой — наоборот.

1b. Вывод величины замедления полимеразы вторичной структурой, образующейся на участке мРНК между рибосомой и полимеразой.

Шпилькой называется цепочка пар спаренных отрезков, которые линейно расположены в петлях друг друга (т.е. соответствующее дерево линейное) с *небольшими* выпячиваниями между соседними парами отрезков и *произвольной* петлей на конце этой цепочки пар; первая пара отрезков называется *черенком* шпильки; как и выше, определяются *плечи* шпильки и *концы* A, B, C, D. В шпильке каждая пара спаренных отрезков, т.е. какая-то гипоспираль, имеет свою петлю, включающую все последующие пары таких отрезков, выпячивания и петли.

Вероятность терминации в зависимости от длины шпильки терминатора (экспериментальные данные из [7]-[9]) имеет вид кривой, которая в физической литературе называется "резонансной". Не претендуя на обсуждение физического процесса взаимодействия шпильки с полимеразой, мы попытались использовать такую кривую для описания зависимости константы скорости перескока полимеразы от вторичной структуры РНК. "Сила" F замедления шпилькой полимеразы, имеющая смысл величины эффективного уменьшения константы скорости движения полимеразы по цепи ДНК, измеряется в с^{-1} и определяется по формуле (9), т.е. считается зависящей только от "частоты" ω:

$$F = \frac{\delta}{L^2(\omega - \omega_0)^2 + 1} \cdot \exp(-\frac{r}{r_0}), \tag{9}$$

где r — расстояние от конца шпильки D до начала полимеразы, биологический смысл параметров L, ω_0, r_0, δ обсуждается ниже, частота ω зависит от всей данной шпильки или, как мы увидим дальше, от макросостояния. Сила воздействия вторичной структуры из несколь-

ких шпилек на полимеразу вычисляется как сумма указанных сил от каждой составляющей структуры шпильки, т.е. считается, что сила действует аддитивно. Таким образом, константа скорости перехода полимеразы с нуклеотида, *не принадлежащего Т-богатому участку*, на следующий нуклеотид определяется по формуле:

$$v(\Omega) = \bar{\lambda}_{pol} - F(\Omega).$$

Перейдем к более удобным переменным, *положив* $\omega = p \cdot \sigma$ и $\omega_0 = p_0 \cdot \sigma$, где p называется *волновым числом* (и, к слову, $p = \frac{2\pi}{\lambda}$, где λ — *длина волны*), а σ — *скорость распространения* колебаний, значение которой не будет нами использоваться. Поэтому перепишем (9) в виде

$$F = \frac{\delta}{L_1^2 \cdot (p - p_0)^2 + 1} \cdot \exp(-\frac{r}{r_0}), \tag{10}$$

где $L_1 = L \cdot \sigma$, а p зависит от шпильки или от макросостояния, которые "замедляют" движение полимеразы. В расчетах везде используется зависимость $F(p)$ вместе с параметрами L_1, p_0, r_0, δ.

В разделе 2 обсуждается биологический смысл всех перечисленных параметров и вычисление значений L_1, p_0, r_0, δ.

Коснемся одной особенности формулы (9), позволяющей вычислять значения L_1 и p_0, зная r_0 и δ, всего лишь по двум точкам. Точнее, пусть при каком-то одном значении r измерено $F_{\mathrm{m}}(r) = \delta \cdot \exp(-\frac{r}{r_0})$ при уже правильных значениях r_0, δ. Пусть $p = p(h, l)$ — корень уравнения $\mathrm{tg}(p \cdot h) = \frac{2}{p \cdot l}$ на интервале от 0 до $\frac{\pi}{2}$, см. формулу (14) ниже. Подставляя $p = p(h, l)$ в исходное F, получим $F(h, l) = \frac{F_{\mathrm{m}}(r)}{1 + L_1^2 \cdot (p - p_0)^2}$ как функцию от h и l. Пусть экспериментально найдены два значения аргументов (h_1, l_1) и (h_2, l_2), для которых: $p(h_1, l_1) > p(h_2, l_2)$ и $F(h_1, l_1) = F(h_2, l_2) = \frac{F_{\mathrm{m}}}{2}$. Тогда искомые L_1 и p_0 легко определяются как

$$\frac{1}{L_1} = \frac{1}{2}\left(p(h_1, l_1) - p(h_2, l_2)\right), \; p_0 = \frac{1}{2}\left(p(h_1, l_1) + p(h_2, l_2)\right). \tag{11}$$

Отсюда получаем, что, зная из эксперимента функцию $F(h, l)$ для двух (подходящих) значений аргументов, можно найти L_1 и p_0 (и отсюда L и ω_0) по формуле (11). В разделе 2 применен в сущности тот же прием, только там длина петли l считается пренебрежимо малой и варьируется длина черенка h — там шпилька состоит только из черенка и петли. В результате получается функция от одной переменной h, у которой легко найти точку $\frac{F_{\mathrm{m}}}{2}$.

Рассмотрим принципиальную *задачу вычисления величины* p, которая зависит от замедляющей шпильки или от замедляющего макросостояния. Это делается за несколько шагов, заголовок к каждому шагу выделен полужирным шрифтом.

Сначала рассмотрим случай **шпильки, состоящей только из черенка и петли**.

Пусть u_0, u_1, $u_2, \ldots, u_n$, u_{n+1} — текущие поперечные координаты $n + 1$ спаренной пары нуклеотидов в черенке, m — масса одной пары нуклеотидов, а $M = \frac{l \cdot m}{2}$ — масса петли, где l — длина петли. Считаем, что связь соседних пар нуклеотидов — упругая с коэффициентом Гука k, и такова же связь последней пары нуклеотидов с петлей, которая считается материальной точкой с массой M. Пусть левый конец этого "камертона" закреплен, т.е. не зависит от времени, $u_0 = 0$. Тогда из второго закона Ньютона имеем

$$m \cdot \ddot{u}_i = -k(u_i - u_{i-1}) - k(u_i - u_{i+1}), \text{ где } i = 1, 2, \ldots, n, \text{ и } M \cdot \ddot{u}_{n+1} = -k'(u_{n+1} - u_n).$$

Считаем, что расстояние между соседними нуклеотидами равно d (малый параметр), и заменяя разности на производные, имеем для функции $u(x, t) = u_i(t)$, $i = \left[\frac{x}{d}\right]$ уравнение в частных

производных

$$m\frac{\partial^2 u}{\partial t^2} = kd^2\frac{\partial^2 u}{\partial x^2},\ u(0,t) = 0,\ \text{и} \tag{12}$$

$$M\frac{\partial^2 u(h,t)}{\partial t^2} = -k'd\frac{\partial u(h,t)}{\partial x}, \tag{13}$$

где $h = nd$ — длина черенка. Заметим, что здесь могут быть использованы и другие линейные и даже нелинейные уравнения; вопрос о правильном выборе уравнения требует отдельного обсуждения.

Положим $c^2 = \frac{kd^2}{m}$, тогда уравнение (12) примет вид $\frac{\partial^2 u}{\partial t^2} = c^2\frac{\partial^2 u}{\partial x^2}$, $u(0,t) = 0$, и это — волновое уравнение, где c — скорость распространения волны.

Его решение ищем в виде

$$u(x,t) = \cos(\omega t + \varphi)\cdot\sin(px),$$

где φ — произвольная начальная фаза колебаний, а функция sin выбрана из условия $u(0,t) = 0$. Из (12) получаем $\omega^2 = p^2c^2$, т.е. $\omega = pc$. Из (13) имеем

$$\omega^2 M\cdot\sin(ph) = k'pd\cdot\cos(ph),$$

т.е.

$$\operatorname{tg}(ph) = \frac{k'pd}{\omega^2 M} = \frac{k'd}{p\cdot Mc^2} = \frac{m}{\beta pMd} = \frac{2}{\beta pl},$$

где $\beta = \frac{k}{k'}$. Здесь и далее положено $d = 1$.

Итак, p находится из уравнения

$$\operatorname{tg}(p\cdot h) = \frac{2}{p\cdot l},\ 0 \le p < \frac{\pi}{2}, \tag{14}$$

а затем ω определяется по формуле $\omega = pc$, где c — скорость распространения колебаний по черенку, h — *длина черенка*, т.е. число спаренных нуклеотидов в нем.

Теперь рассмотрим шпильку, состоящую из нескольких спаренных **отрезков с небольшими выпячиваниями** между ними и произвольной петлей на конце.

Сначала предположим, что имеется ровно **два таких отрезка** с длинами h_1 и h_2 и выпячиванием между ними с массой M_1, и петлей на конце с массой M_2. Выбирая единицы измерения так, чтобы $c = 1$, $m = d$, $k = 1/d$, запишем волновое уравнение $\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}$ при $0 < x < h_1$ и $h_1 < x < h_1 + h_2 = h$ (где h — суммарная длина двух отрезков), с краевыми условиями $u(0,t) = 0$, $M_2\frac{\partial^2 u(h,t)}{\partial t^2} = -\frac{\partial u(h,t)}{\partial x}$, а также с условиями $M_1\frac{\partial^2 u(h_1,t)}{\partial t^2} = \frac{\partial u(h_1+0,t)}{\partial x} - \frac{\partial u(h_1-0,t)}{\partial x}$, $u(h_1+0,t) = u(h_1-0,t)$. Решение ищем в виде

$$u(x,t) = \cos(\omega t + \varphi)\cdot f(x),$$

где $f(x) = \sin(px)$ при $0 < x < h_1$ и $f(x) = A\sin(py) + B\cos(py)$, $y = x - h_1$ при $0 < y < h_2$.

Отсюда получаем

$$\cos(ph) - pM_1\sin(ph_1)\cdot\cos(ph_2) = pM_2(\sin(ph) - pM_1\sin(ph_1)\sin(ph_2)). \tag{15}$$

Предполагая, что M_1 мало, положим $p = \bar{p} + z$, где $\bar{p}$ удовлетворяет уравнению

$$\operatorname{tg}(\bar{p}h) = \frac{1}{\bar{p}M_2}.$$

В первом порядке по M_1 получим

$$z = -\bar{p} \cdot \frac{M_1 \sin^2 \bar{p}h_1}{h + M_2 \sin^2 \bar{p}h},$$

т.е.

$$p = \bar{p} \cdot \left(1 - \frac{M_1 \sin^2 \bar{p}h_1}{h + M_2 \sin^2 \bar{p}h}\right).$$

Восстанавливая константу $\rho = \frac{m}{d}$, получим

$$p = \bar{p} \cdot \left(1 - \frac{M_1 \sin^2 \bar{p}h_1}{\rho h + M_2 \sin^2 \bar{p}h}\right) = \bar{p} \cdot \left(1 - \frac{l_1 \sin^2 \bar{p}h_1}{2h + l \sin^2 \bar{p}h}\right), \tag{16}$$

где $\bar{p}$ находится из уравнения $\operatorname{tg}(\bar{p}h) = \frac{\rho}{\bar{p}M_2} = \frac{2}{\bar{p}l}$, т.е.

$$\operatorname{tg}(\bar{p}h) = \frac{2}{\bar{p}l}. \tag{17}$$

Далее частота находится по формуле $\omega = pc$.

Теперь рассмотрим случай, когда шпилька содержит s **отрезков с длинами** $h_1, \ldots, h_s$ и $s-1$ выпячиваний между ними с массами $M_1, \ldots, M_{s-1}$ и длинами $l_1, \ldots, l_{s-1}$ и, наконец, петлю с массой M и длиной l, предполагая **малость** этих выпячиваний, но не самой петли. Тогда аналогично получим

$$p = \bar{p} \cdot \left(1 - \frac{1}{\rho \cdot h + M_s \cdot \sin^2(\bar{p} \cdot h)} \cdot \sum_{i=1}^{s-1} M_i \cdot \sin^2(\bar{p} \cdot h(i))\right),$$

$$p = \bar{p} \cdot \left(1 - \frac{1}{2h + l \cdot \sin^2(\bar{p} \cdot h)} \cdot \sum_{i=1}^{s-1} l_i \cdot \sin^2(\bar{p} \cdot h(i))\right), \tag{18}$$

где $h(i) = h_1 + \ldots + h_i$, $h = h(n) = h_1 + \ldots + h_n$ и $\bar{p}$ находится из аналогичного уравнения $\operatorname{tg}(\bar{p}h) = \frac{\rho}{\bar{p} \cdot M_s}$, т.е. из уравнения

$$\operatorname{tg}(\bar{p} \cdot h) = \frac{2}{\bar{p} \cdot l}, \tag{19}$$

где $\rho = m$ с учетом того, что принято $d = 1$, а в (10) положили $m = 1$.

Случай макросостояния. Теперь определим частоту ω по макросостоянию Ω, сведя этот вопрос к некоторому микросостоянию ω'. Макросостояние Ω (которое можно называть также вторичной структурой на участке между рибосомой и полимеразой) однозначно разлагается в цепочку *"неразложимых макросостояний"*, каждое из которых определяется наличием *черенка макросостояния* — самой внешней скобки макросостояния. По этой скобке однозначно определяется *пара* A и D — внешних концов самой внешней (непродолжаемой) спирали этого неразложимого макросостояния.

По неразложимому макросостоянию мы хотим определить "силу" $F(\Omega)$, основываясь на формуле (10), а затем суммировать по всем неразложимым макросостояниям. Трудность состоит в том, что неразложимое макросостояние определяет семейство весьма различных реализующих его шпилек и других микросостояний, и, естественно, в самом макросостоянии не даются, например, значения величин h, $h(i)$, l, l_i. Эта трудность может преодолеваться несколькими способами. Укажем один из них.

По каждому микросостоянию c, реализующему данное неразложимое макросостояние Ω, определим, как указано ниже, одну конкретную шпильку ω' — *корень микросостояния* ω и затем вычислим $F(\Omega)$ как математическое ожидание по всем микросостояниям $p = p(c)$, реализующим данное Ω, применяя формулы (18-19):

$$F(\Omega) = \sum_{\omega\in\Omega} p(\omega) \cdot \sum_i F(\omega'_i), \tag{20}$$

где суммирование по i соответствует всем неразложимым микросостояниям (т.е. всем $\langle A_i, D_i\rangle$) данного ω. Итак, ω'_i начинается с гипоспирали, соответствующей какой-то паре нуклеотидов $\langle A_i, D_i\rangle$, и продолжается по участку A_i, D_i исходной последовательности в соответствии со спариваниями, фиксированными в ω, игнорируя мелкие выпячивания в ω до появления в нем большого выпячивания (по некоторому порогу) или разветвления. Участки до этого момента считаются *плечами корня* ω', а участок, остающийся с этого момента, объявляется *петлей корня* ω'. Затем к корню ω', являющемуся, таким образом, некоторой шпилькой, применяются формулы (18-19). Здесь возможны варианты, связанные с учетом разветвлений, слишком большой петли, с учетом массы всей шпильки, и т.п.

1c. Движение полимеразы по цепи ДНК.

Опишем перескок полимеразы с нуклеотида, *принадлежащего T-богатому участку*. Если конец полимеразы, обозначаемый z, находится на n-м нуклеотиде, то возможен ее перескок на $(n+1)$-й нуклеотид или срыв с нуклеотидной последовательности. Возможны три варианта организации этого перескока или срыва, причем они в модели дают численно одинаковые результаты.

Первый вариант. Полимераза сначала переходит на *условный* $\left(n+\frac{1}{2}\right)$-й нуклеотид с константой скорости перехода $\bar\lambda_{pol} = 40\text{с}^{-1}$, а затем может сорваться с константой скорости, равной $\lambda_{ur} \sim 10\text{с}^{-1}$, если z находится на нуклеотиде из Т-богатого участка, и равной 0 в ином случае. Или может перейти на $(n+1)$-й нуклеотид с константой скорости

$$\nu(\Omega) = \frac{(\bar\lambda_{pol})^2}{F(\Omega)} - \bar\lambda_{pol}, \tag{21}$$

где Ω — макросостояние на участке между концом x рибосомы и началом y полимеразы, т.е. на участке $[x, y]$. Если макросостояние Ω пустое, то $F(\Omega) = 0$ и $\nu(\Omega) = \infty$.

T-богатый участок определяется следующим образом. Нуклеотид z назовем Т-богатым, если существует хотя бы одно слово, содержащее z на любом его месте, которое по длине больше порога (например, 6) и по плотности нуклеотидов T больше порога (например, 0.8). Это слово может содержать исключения, т.е. не букву T, где угодно, включая и концы; само z также может быть не T. В множестве всех Т-богатых нуклеотидов образуем все максимальной длины интервалы. Они называются Т-богатыми участками, и не пересекаются.

Алгоритмически эти участки ищутся заранее для всей исходной последовательности очевидным образом: для каждой длины d, начиная от 6, вдоль последовательности сдвигается интервал длины d. Плотность каждого следующего (сдвинутого на 1 нуклеотид) интервала считается по плотности предыдущего: если "новая" буква есть T, а "оставленная" не T, то плотность увеличивается на $1/d$, если наоборот, то уменьшается на $1/d$, а иначе не меняется.

Каждый раз, когда плотность больше порога, все ещё не помеченные нуклеотиды интервала помечаются как T-богатые.

В частности, поле урацилов является T-богатым участком. Говоря о букве T, конечно, имеют в виду букву A на комплементарной цепи. Случай нескольких T-богатых участков в исходной последовательности требует специального учета.

Формула (21) основана на следующем соображении. Принимается, что среднее время двух переходов из n в $n+\frac{1}{2}$ и затем в $n+1$ равно $\frac{1}{\bar{\lambda}_{pol}-F(\Omega)}$. С другой стороны, это время равно $\frac{1}{\bar{\lambda}_{pol}}+\frac{1}{\nu(\Omega)}$. Поэтому

$$\frac{1}{\bar{\lambda}_{pol}}+\frac{1}{\nu(\Omega)}=\frac{1}{\bar{\lambda}_{pol}-F(\Omega)}, \tag{22}$$

откуда получаем (21).

Если шпилька состоит из одного черенка с пренебрежимо малой петлей, то $p=\frac{\pi}{2h}$ и $F=\frac{\delta}{L_2^2\cdot\left(\frac{1}{h}-\frac{1}{h_0}\right)^2+1}\cdot\exp(-\frac{r}{r_0})$, где $L_2=\frac{\pi}{2}\cdot L_1$, $h_0=\frac{\pi}{2p_0}$. Вероятность полимеразе перейти с n-го нуклеотида на $(n+1)$-й и не сорваться, очевидно, равна

$$\frac{\nu}{\nu+\lambda_{ur}}, \tag{23}$$

таким образом, она отлична от 1 только на T-богатом участке и только при наличии шпилек (т.е. когда $\nu<\infty$ или то же $F>0$). Вероятность совершить N "двойных" переходов из n в $n+\frac{1}{2}$ и сразу затем в $n+1$ на T-богатом участке равна, очевидно,

$$\left(\frac{\nu}{\nu+\lambda_{ur}}\right)^N. \tag{24}$$

Вероятность полимеразе пройти T-богатый участок, расположенный между r_{min} и r_{max} (расстояние измеряется от конца фиксированной шпильки до начала и конца этого участка), равна $\Pr=\prod\limits_{i=r_{min}}^{r_{max}}\frac{\nu_r}{\varepsilon+\nu_r}$, $\Pr^{-1}=\prod\limits_{i=r_{min}}^{r_{max}}(1+\frac{\varepsilon}{\nu_r})$, $\frac{\varepsilon}{\nu_r}=\varepsilon(\frac{1}{\bar{\lambda}_{pol}-F(r)}-\frac{1}{\bar{\lambda}_{pol}})=\frac{\varepsilon\cdot F(r)}{\bar{\lambda}_{pol}\cdot(\bar{\lambda}_{pol}-F(r))}$, где $F(r)$ — сила, замедляющая полимеразу при данном значении r.

Например, формулу (22) можно применить к следующим известным из эксперимента данным о зависимости частоты терминации от длины черенка (т.е. от числа пар h нуклеотидов в нем при поле урацилов длины 8, тогда $N=7$) [7]-[9]:

$$\langle 3,0.2\rangle,\ \langle 7,0.8\rangle,\ \langle 14,0.2\rangle.$$

В этих парах первое число указывает на длину черенка, а второе — на частоту терминации. А именно, функция F в приближении, когда она записывается как функция от h, зависит от трех параметров h_0, L_2, δ, и получается система трех нелинейных уравнений с тремя неизвестными. Решая ее приблизительно и округляя, получим:

$$h_0=7, \delta=25,\ L_2=22{-}23,\ L_1=14.5 \text{ (при ширине резонанса 5)},\ p_0=\frac{\pi}{14}. \tag{25}$$

Эти численные значения являются ориентировочными и требуют уточнения, в том числе, с учетом филогенетической группы организма. Хотя формула (11) и еще способ, указанный в разделе 2, позволяют более обоснованно определять значения этих параметров, они требуют большего объема экспериментальных данных, чем указанные выше три точки.

Второй вариант. Кроме схемы, указанной выше — ее содержание отражено в соотношении (22), — можно предложить такую схему, эквивалентную первой в смысле времени прохождения (22) и вероятности прохождения (23). Полимераза из положения $z = n$ может перейти в положение $z = n + 1$ с константой скорости $\bar{\lambda}_{pol} - F(\Omega)$ и сорваться с константой скорости μ, где $\frac{\mu}{\bar{\lambda}_{pol}-F} = \frac{\lambda_{ur}}{\nu}$, т.е.

$$\mu = (\bar{\lambda}_{pol} - F) \cdot \frac{\lambda_{ur}}{\nu} = (\bar{\lambda}_{pol} - F) \cdot \lambda_{ur} \cdot (\frac{1}{\bar{\lambda}_{pol} - F} - \frac{1}{\bar{\lambda}_{pol}}) = \frac{\lambda_{ur} F}{\bar{\lambda}_{pol}}.$$

Откуда

$$\mu = \frac{\lambda_{ur} F}{\bar{\lambda}_{pol}}. \tag{26}$$

Здесь можно оценить $\frac{\bar{\lambda}_{pol}}{\lambda_{ur}} = 4$. Это отношение является естественным параметром модели.

Третий вариант. Еще одна схема могла бы быть такой: полимераза из положения $z = n$ может перейти в положение $z = n+1$ с константой скорости $\bar{\lambda}_{pol}$ и в состояние n^*, из которого может соскочить с константой λ_{ur} или вернуться назад в $z = n$. Если переходы между n ("основное состояние") и n^* ("возбужденное состояние") быстрые, то эту схему можно заменить ее усреднением: переход из n в $n+1$ с константой $\alpha \cdot \bar{\lambda}_{pol}$ и срыв из n с константой $(1-\alpha) \cdot \lambda_{ur}$, где α — вероятность найти полимеразу в основном состоянии, а $(1 - \alpha)$ — вероятность найти ее в возбужденном состоянии. Приравнивая $\alpha \cdot \bar{\lambda}_{pol} = \bar{\lambda}_{pol} - F$, получим $(1 - \alpha) \cdot \bar{\lambda}_{pol} = F$, т.е. $(1 - \alpha) = \frac{F}{\bar{\lambda}_{pol}}$, и окончательно имеем для перескока полимеразы константу $\bar{\lambda}_{pol} - F(\Omega)$ и для срыва полимеразы константу

$$\mu = \frac{\lambda_{ur} F}{\bar{\lambda}_{pol}}. \tag{27}$$

Таким образом, все три варианта (три схемы) в части параметров, отслеживаемых в модели, дают одинаковые результаты.

1d. Движение рибосомы по цепи мРНК.

На *нерегуляторных* кодонах константа скорости сдвига рибосомы на 1 кодон принимается равной $\bar{\lambda}_{rib} = 15\text{с}^{-1}$. На *регуляторных* кодонах λ_{rib} зависит от концентрации c соответствующей аминокислоты по формуле:

$$\lambda_{rib}(c) = \bar{\lambda}_{rib} \cdot \left(1 - \exp\left(-\frac{c}{c_0}\right)\right),$$

или (как нами реально принималось) по формуле Микаэлиса-Ментен:

$$\lambda_{rib}(c) = \frac{\bar{\lambda}_{rib} \cdot c}{c_0 + c}, \tag{28}$$

где c_0 — концентрация заряженных тРНК, при которой рибосома движется по регуляторным кодонам со скоростью равной половине от максимальной скорости такого движения $\bar{\lambda}_{rib} = 15$.

В связи с тем, что неясно, как экспериментально измерять такое c_0 (а модель должна ориентироваться на сравнение с результатами экспериментов), был принят следующий подход. Зависимость концентрации заряженных тРНК от концентрации аминокислоты определяется также по формуле Микаэлиса-Ментен. Результат ее подстановки в формулу (28) снова приводит к формуле такого же вида, в которой теперь c — концентрация аминокислоты в клетке. Но в экспериментах концентрация фиксируется не в клетке, а вне нее, в культуре. Поэтому, подставляя еще раз, получим ту же формулу (28), где окончательно c — *концентрация*

аминокислоты в культуре, соответствующее c_0 — параметр Микаэлиса-Ментен, отражающий эти два процесса: влияние концентрации аминокислоты в культуре на ее же концентрацию в клетке, влияние концентрации в клетке на концентрацию заряженных тРНК, а ее — на вероятность движения рибосомы на регуляторных кодонах. Таким образом, эта константа c_0 не имеет прямой биологической интерпретации.

1e. Посадка свободной рибосомы на область Шайн-Дальгарно.

Как только область Шайн-Дальгарно и старт-кодон (atg или gtg) лидерного пептида транскрибированы, появляется возможность свободной рибосоме связаться с мРНК. Представим себе (речь идет о рассуждении, которое лишь поможет нам вывести формулу), что комплекс "рибосома и загруженная тРНК" являются двумя "плечами", которые должны связаться соответственно с областью ШД и старт-кодоном, при этом на участке мРНК, включающем ШД и старт-кодон, имеется некоторое макросостояние Ω. Возникает возможность переходов между состояниями $\langle \Omega, freerib \rangle$ и $\langle \Omega, boundrib \rangle$. Константы скоростей переходов между ними обозначим соответственно слева направо K_{in} и наоборот K_{out}. По общему правилу

$$K_{in} = \sum_{\omega \in \Omega} \sum_{\omega' \in \langle \Omega, boundrib \rangle} p(\omega) \cdot K(\omega \to \omega'), \tag{29}$$

где $K(\omega \to \omega') = \kappa_{SD} \cdot \exp\left(G_{loop}(\omega) - G_{loop}(\omega')\right)$ и $\kappa_{SD} = 10\text{с}^{-1}$. По-видимому, можно считать, что $G_{loop}(\omega) = G_{loop}(\omega')$, и получим $K(\omega \to \omega') = \kappa_{SD}$ и

$$K_{in} = \kappa_{SD} \cdot card\{\omega' | \omega' \in \langle \Omega, boundrib \rangle\}. \tag{30}$$

Обратный переход имеет константу скорости

$$K_{out} = \sum_{\omega' \in \langle \Omega, boundrib \rangle} \sum_{\omega \in \Omega} p(\omega') \cdot K(\omega' \to \omega), \tag{31}$$

где $K(\omega' \to \omega) = \kappa \cdot \exp\left(G_{hel}(\omega') - G_{hel}(\omega)\right)$ и $\kappa = 10^6\text{с}^{-1}$ — стандартная константа скорости замыкания. В модели рассматривался также симметричный вариант, в котором

$$K_{in} = \kappa_{SD} \text{ и } K_{out} = \sum_{\omega' \in \langle \Omega, boundrib \rangle} \sum_{\omega \in \Omega} \frac{p(\omega') \cdot K(\omega' \to \omega)}{\text{card}\langle \Omega, boundrib \rangle}, \tag{32}$$

где $K(\omega' \to \omega) = \kappa \cdot \exp\left(G_{hel}(\omega') - G_{hel}(\omega)\right)$.

Из состояния $\langle \Omega, boundrib \rangle$ возможны переход в состояние $\langle \Omega, freerib \rangle$ или сдвиг рибосомы на 1-й (после старта) кодон, в результате чего рибосома уже не может стать свободной (до достижения стоп-кодона ЛП, когда рибосома мгновенно "срывается"), инициация заканчивается и рассматриваются переходы, характерные для установившегося движения, когда рибосома и полимераза уже обе находятся на цепи мРНК.

1f. Учет белок-ДНКового взаимодействия при посадке и срыве полимеразы.

Моделирование без учета репрессии и активации посадки полимеразы (и фактора срыва полимеразы, например, с участием ТРАП-подобного белка) может приводить к плохому согласованию с экспериментальными данными. Поэтому по аналогии с регуляцией триптофана у *E. coli* предполагается наличие белок-ДНКовой регуляции, правда, пока в очень упрощенном виде. А именно, пусть $v(c)$ — зависимость числа полимераз, садящихся в единицу времени на промотор, от той же *концентрации* c аминокислоты. Тогда $v(c) \cdot (1 - p(c))$ дает число транскрипций в единицу времени. Иными словами, $1 - p(c)$ — это вероятность антитерминации при условии, что полимераза уже села, а $v(c)$ — вероятность посадки полимеразы. В текущей

модели принимается, что $\nu(c)$ в относительных единицах является ступенчатой функцией, равной 0 при концентрации равной 0. Такая функция задается указанием начала c_i и высоты d_i каждой ступеньки. Альтернативные формулы $\nu(c) = \exp\left(-\frac{c}{c_0}\right)$ или $\nu(c) = \frac{c_0}{c_0+c}$, по-видимому, приводят к худшему согласованию с экспериментом.

2. СРАВНЕНИЕ РЕЗУЛЬТАТА С ЭКСПЕРИМЕНТАЛЬНЫМИ ДАННЫМИ. ВЫЧИСЛЕНИЕ ПАРАМЕТРОВ СИЛЫ ЗАМЕДЛЕНИЯ. ОПИСАНИЕ СХЕМЫ МОДЕЛИРОВАНИЯ

Цель и сравнение. В случае классической аттенюаторной регуляции *цель моделирования* состояла в численном определении зависимости $p = p(c)$ вероятности терминации от концентрации c заряженных тРНК (или: от концентрации c аминокислоты в клетке, или: от концентрации c аминокислоты вне клетке, в культуре) для различных биологических регуляторных областей, а также ряда связанных с $p(c)$ зависимостей. Более подробному учету одновременной белок-ДНКовой регуляции того же гена предполагается посвятить отдельную публикацию. Конечно, три обозначенные выше зависимости различны.

Для построения зависимости $p = p(c)$ при каждом значении c из сетки с некоторым шагом узлов указанный в нашей модели процесс проигрывался определенное число N раз (например, $N = 10^3$-10^4 раз, что дает примерно одинаковый результат) и вычислялось $p = p(c)$ как доля случаев, в которых происходила терминация.

Из экспериментов известны значения, с которыми могут сравниваться результаты моделирования: например, отношение вероятностей $p = p(c)$ при достаточно большой и достаточно малой концентрациях. Известны графики зависимости активности фермента (например, количество атранилат синтазы AS в наномолях, деленное на количество всех белков в миллиграммах) от концентрации аминокислоты (например, триптофана в культуре в микромолях на литр [10]). Чтобы результаты моделирования можно было сравнить с этими экспериментальными данными, они интерпретируются как третья из упомянутых выше зависимостей. Однако и в этом случае переход к физическим единицам измерения активности и концентрации от условных единиц на осях графика функции $1 - p(c)$ (даже с учетом репрессии) представляется нетривиальным. Предполагается получать согласование этих графиков и значение параметра c_0 для одних регуляторных областей и затем проверять их на родственных случаях.

Вычисление параметров силы замедления. Параметрами функции замедления F, см. (10), являются: $\bar{\lambda}_{pol} = 40\text{с}^{-1}$ — константа скорости перескока полимеразы в отсутствии замедления, r_0 — расстояние, на котором шпилька снижает свой воздействие в e раз по сравнению с силой из (10), взятой на нулевом расстоянии, δ — коэффициент влияния шпильки (может быть, зависящий от доли пар gc в ее составе), ω_0 — точка максимума функции $F(\omega)$, L и L_1 — (по сути один и тот же) параметр, характеризующий "широту резонанса" между шпилькой и полимеразой. Итак, сила F зависит от четырех параметров r_0, δ, ω_0, L_1, постоянных для генома. Значение ω зависит от конкретных особенностей шпилек, составляющих Ω.

Рассмотрим еще раз вопрос о численном определении этих четырех параметров, определяющих "силу" F, который в конечном счете сводится к объему доступных результатов экспериментов. О величине параметра r_0 можно судить по типичной длине поля остатков урацила, например, $r_0 \approx 1$–5.

Параметр δ показывает, насколько велико максимально возможное замедление полимеразы по сравнению с его отсутствием. Например, если максимально возможное замедление в 2 раза больше по сравнению с его отсутствием, то $\delta = 20$. По-видимому, это замедление может достигать значений в 2-5 раз.

Параметр ω_0 — это значение частоты ω, при котором $F(\omega)$ достигает максимума, или, тоже самое, p_0 — это значение волнового числа, при котором $F(p)$ достигает максимума.

Рассмотрим несколько другой, чем формула (11), способ определения значений параметров p_0 и L_1. *Ширина резонанса* определяется как число $wr = b - a$, где числа a и b определяются из условия $F(a) = F(b) = \frac{F_{max}}{2}$ и $F_{max} = F(p_0)$ — максимум функции $F(p)$. Легко выразить значение L_1 через ширину резонанса как $L_1 = \frac{2}{wr}$. Для шпильки, состоящей только из черенка с пренебрежимо малой петлей, зависимость $p(h)$, где h — число пар спаренных нуклеотидов в черенке, по формуле (14) имеет вид

$$p \cdot h = \frac{\pi}{2}.$$

Пусть p_0 связано с некоторым новым h_0 *соотношением* $p_0 \cdot h_0 = \frac{\pi}{2}$. Перейдя к переменной h, найдем L_1 по, как можно думать, *известной из эксперимента* функции $F(h)$. А именно, $L_1 = \frac{2h_0 \cdot (h_0+\delta)}{\pi \cdot \delta}$, где δ — *правая полуширина* резонанса функции $F(h)$, но лучше использовать формулу $L_1 = \frac{2h_0^2}{\pi} \cdot \left(\frac{1}{sr} + \sqrt{\frac{1}{sr^2} - \frac{1}{h_0^2}}\right)$, где sr — ширина резонанса функции $F(h)$. По функции $F(h)$ находится и h_0 как точка ее максимума (конечно, $F(h)$ не симметрично относительно h_0) или лучше по формуле $h_0^{-1} = \frac{1}{2}(h_1^{-1} + h_2^{-1})$, где h_1 и h_2 — концы интервала, на котором максимальное значение $F(h)$ убывает в два раза. Итак, на интервале $\left[h_0 - \frac{h_0^2}{\frac{\pi}{2} \cdot L_1 - h_0},\ h_0 + \frac{h_0^2}{\frac{\pi}{2} \cdot L_1 + h_0}\right]$ влияние вторичной структуры ослабевает примерно в 2 раза по сравнению с ее влиянием при $h = h_0$. Длина этого интервала равна $\frac{\pi \cdot L \cdot h_0^2}{\left(\frac{\pi}{2} \cdot L\right)^2 - h_0^2}$.

Описание схемы моделирования. Для исходной фиксированной последовательности РНК *текущее состояние* характеризуется:

1) *Окном* между положениями конца x рибосомы и начала y полимеразы; "размер" рибосомы от ее активного центра обозначим s_0 (порядка 10-12), а "размер" полимеразы от y — места выхода РНКовой цепи до точки расхождения ДНК и РНК через s_1 (порядка 1-5 нуклеотида); точку расхождения цепей ДНК и РНК, в которой она "стоит" на определенном нуклеотиде обозначаем z. В *окне* происходит перестройка вторичной структуры от одного макросостояния Ω к другому Ω' — при этом макросостояния могут включать только спирали, пересекающиеся с окном обоими плечами хотя бы по одному нуклеотиду, т.е. речь идет о *макросостояниях в окне* (для данного окна).

2) *Списком* T (потенциальных) спиралей, пересекающихся с окном обоими плечами (*хотя бы по* 1 нуклеотиду); это тривиальная компонента состояния в том смысле, что можно каждый раз вычислять ее по исходному списку спиралей.

3) *Макросостоянием* Ω, оно же: непустая диаграмма, вторичная структура в окне.

До посадки полимеразы окна нет (*пустое окно*), а после посадки полимеразы и до посадки рибосомы окно начинается в первом нуклеотиде исходной последовательности — точке 0 и заканчивается в текущем положении начала полимеразы. В окне может впервые появиться непустое макросостояние Ω, состоящее из одной хорды. Затем к этой хорде может добавиться вторая хорда или, наоборот, макросостояние может вернуться к исходному — пустому.

Отслеживается один из *двух возможных исходов* моделирования: 1) событие срыва полимеразы на одном из нуклеотидов поля урацилов исходной последовательности, или 2) полимераза проходит все поля урацилов. Обычно конец поля урацилов совпадает с концом исходной биологической последовательности.

Инициация процесса РНКовой терминации: от посадки полимеразы до посадки рибосомы.

1) Полимераза садится на промотор и через некоторое число шагов приходит в точку старта лидерного пептида по общему правилу.

2) Как только полимераза прошла старт лидерного пептида и еще s_0 нуклеотидов, на область ШД пытается сесть рибосома с константой скорости, которая отражает зависимость от качества этой области и от вторичной структуры, закрывающей ее. Как только это произошло, активный центр рибосомы занимает положение на старте ЛП. В этот момент фиксируются: левый конец x окна в точке "старт лидерного пептида"$+s_0$, и правый конец y окна в том положении, которое на тот момент занимает начало полимеразы. Всегда выполняется $z = y + s_1$.

Переходы в процессе РНКовой терминации после формирования окна $[x, y]$.

1) *Сдвиг* полимеразы на 1 нуклеотид вправо, при этом окно увеличивается на 1 нуклеотид, а список спиралей T, вообще говоря, расширяется. Или *срыв* полимеразы на Т-богатом участке.

2) *Сдвиг* рибосомы на 1 кодон (3 нуклеотида) вправо; окно уменьшается на 3 нуклеотида и, вообще говоря, список спиралей T сокращается, а макросостояние Ω меняется. Из диаграммы Ω исключается *самая левая* скобка, если приписанная ей спираль не входит в новый список T (и, конечно, соответствующая правая скобка). Так полученное макросостояние – новое Ω, может быть, пустое — фиксируется в текущем окне.

Заметим, что здесь не нужно заново проверять непустоту старого макросостояния Ω в новом окне: когда проверялась непустота диаграммы на большем участке, ее непустота (как и всех ее поддиаграмм) проверялась и на меньших участках, и надо лишь помнить соответствующую информацию.

3) *Перестройка* вторичной структуры, т.е. смена макросостояния в окне; при этом само окно и список спиралей T не меняются.

Окончание моделирования.

При наступлении события срыва полимеразы моделирование прекращается; в ином случае полимераза проходит все поле урацилов и моделирование также прекращается. Обычно сама исходная биологическая последовательность заканчивается на конце поля урацилов.

При каждом переходе, если на нем рибосома не сдвигается, то можно фиксировать и время до наступления перехода, эти времена суммируется вплоть до наступления события первого сдвига рибосомы. Если такое суммарное время t превосходит некоторый порог, то рибосома может "разрезать" исходную последовательность.

Аналогично можно подсчитывать суммарное время перекрывания области ШД: например, сколько времени был закрыт каждый из входящих в него нуклеотидов. Вместо порога здесь можно использовать функцию вероятности разрезания вследствие "долгого стояния" t рибосомы.

Организация переходов при моделировании.

Далее моделирование происходит стандартным образом с использованием программы, реализующей метод Монте-Карло, или оригинальной программы из [6]. *Состояние* характеризуется окном с левым x и правым y концами, концом z полимеразы, списком T спиралей в этом окне, макросостоянием Ω в окне. Итак, она характеризуется набором $\langle x, y, z, T, \Omega\rangle$. В период инициации в описание ситуации еще входит признак ς — села рибосома или нет; затем его можно опускать.

Окрестностью данного состояния Ω (с *центром* в Ω) называется набор всех состояний, в которые можно (с ненулевой вероятностью) перейти из Ω. Если окрестность состоит из n состояний и соответствующие константы скоростей переходов равны соответственно $k_1, \ldots, k_n$ (пусть $k = \sum k_i$), то состояние, в которое переходим (которое считается следующим на данной траектории), определяется как реализация случайной величины $i \to \frac{k_i}{k}$. При этом в некоторые моменты моделирования дополнительно определяется время до наступления перехода как реализация случайной величины $t \to k \cdot e^{-kt}$. Заметим, что порядки величин λ_{sd}, λ_{rib}, λ_{pol} и $K(\Omega \to \Omega')$ значительно отличаются.

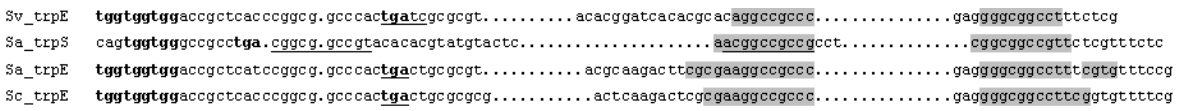

Рис. 1. Выравнивание для 4 стрептомицетов: полужирным шрифтом выделены регуляторные и стоп-кодоны, подчеркиванием — антитерминатор, серым фоном — терминатор

Анализ модели и численный счет показал, что *наиболее критическими параметрами модели* являются: L_1, r_0, α. Ниже результаты счета приведены при различных значениях этих параметров.

3. ЧИСЛЕННЫЕ ЗНАЧЕНИЯ ПАРАМЕТРОВ И ОПИСАНИЕ ПРОГРАММЫ

Исходная последовательность бралась от начала промотора до конца поля урацилов. В качестве примера в следующем разделе 4 будут приведены результаты счета для четырех генов стрептомицетов *S. venezuelae* ISP5230, *S. avermitilis* MA-4680 и *S. coelicolor* A3(2) (далее будем ссылаться на них по именам Sv_trpE, Sa_trpS, Sa_trpE и Sc_trpE), выравнивание которых из [11] приведено на рисунке 1. Массовый счет для других организмов, в том числе для Грам-отрицательных бактерий, будет представлен в последующих публикациях.

Были установлены следующие значения параметров. Универсальная газовая постоянная 8,31 Дж/(моль·°K), температура 310°K (37°C). Значения энергии стекинга для пары спаренных нуклеотидов бралось из таблицы 1, которая относится к температуре 37°C (310°K):

Таблица 1. Энергия стекинга в ккал/моль.

cg	**gc**	**gu**	**ug**	**au**	**ua**	
−2.40	−3.30	−2.10	−1.40	−2.10	−2.10	**cg**
−3.30	−3.40	−2.50	−1.50	−2.20	−2.40	**gc**
−2.10	−2.50	1.30	−0.50	−1.40	−1.30	**gu**
−1.40	−1.50	−0.50	0.30	−0.60	−1.00	**ug**
−2.10	−2.20	−1.40	−0.60	−1.10	−0.90	**au**
−2.10	−2.40	−1.30	−1.00	−0.90	−1.30	**ua**

Полагаем параметры $B = 5$ и $C = 5$. Константа медленного замыкания $\kappa = 10^6 \text{с}^{-1}$. Параметр c_0 — концентрация заряженных тРНК, при которой рибосома движется с "нормативной" скоростью, сначала полагаем равным 1, т.е. измерения по оси c относятся к долям $\frac{c}{c_0}$.

Параметры функции замедления F варьировались, но мы ориентировались на их значения: $h_0 = 7$, $\delta = 25$, $L_2 = 22{-}23$, $L_1 = 14.5$ (при ширине резонанса 5), $p_0 = \frac{\pi}{14}$, $r_0 = 2{-}5$.

"Размер" рибосомы принимается равным $s_0 = 10{-}12$ нуклеотидов: он измеряется от места, где транслируется кодон — ее "активного центра" — до начала окна x, т.е. до места, с которого она расплетает вторичную структуру на РНК. Рибосома пытается сесть на ШД, как только полимераза транскрибировала старт-кодон ЛП и находится на расстоянии s_0 нуклеотидов от старта ЛП. Рибосома (в момент ее посадки на ШД) имеет активный центр на старт-кодоне, т.е. x определяется как "s_0 нуклеотидов после старта ЛП".

"Размер" полимеразы $s_1 = 2{-}7$ нуклеотидов: от выхода РНК из нее, т.е. от конца окна y — начала полимеразы, до транскрибируемого нуклеотида (места, обозначаемого z, разделения ДНК и РНК) — это z проверяется на попадание в Т-богатый участок (именно с такого z полимераза может соскочить). Значение s_1 весьма важно и, к сожалению, заранее не ясно.

Спираль имеет плечо от 3 и петлю от 3 до 40 нуклеотидов. Гипоспираль имеет плечо от 3 нуклеотидов. Для вычисления константы скорости добавления или исчезновения гипоспирали с минимальными длинами используется общая формула, как при добавлении или исчезновении любой хорды микросостояния. В то же время альтернативное рассмотрение произвольных по длине плеча гипоспиралей кажется вполне обоснованным.

Описание программы. Программная реализация модели выполнена на языке C++, текст программы содержит более 2,5 тыс. операторов. Учитывая предполагаемый большой объем вычислений, а в перспективе и массовые вычисления, особое внимание обращалось на возможность использования программы на различных платформах (от ПК до рабочих станций и суперкомпьютеров) и в среде различных операционных систем, прежде всего, Windows и Unix. Поэтому программа выполнена как 32-разрядное консольное приложение RNAmodel с интерфейсом командной строки, в которой могут задаваться значения параметров модели для конкретного счета. Для важнейших параметров предусмотрены значения, принимаемые по умолчанию. Исходные данные и результаты вычислений имеют вид текстовых файлов, имена которых также являются параметрами командной строки. Эти текстовые файлы удобно импортируются в программы для последующей обработки; в частности, с помощью Excel результаты также для большей наглядности могут представляться в графической форме (именно такие графики приводятся ниже для конкретных проведенных вычислений).

Главная особенность реализации состоит в построении нами специальной структуры данных, описывающих текущее состояние модели, рис. 2. Хотя объектная иерархия и не реализуется напрямую в программе, в модели хорошо заметна иерархия “макросостояние-микросостояние-гипоспираль-спираль”. При этом объекты нижнего уровня (потенциальные спирали) являются *статическими* (т.е. не зависят от ширины и положения окна) и потому находятся однократно для всей исходной последовательности и затем хранятся в отдельном массиве, на который вышестоящие уровни будут ссылаться по номеру спирали.

Коль скоро медленные переходы осуществляются между макросостояниями, отличающимися не более чем на одну хорду, совокупности объектов более высоких уровней организованы в виде *разделов* в соответствии с длиной диаграммы макросостояния. Эти совокупности объектов уже являются *динамическими*, т.е. в программе перестраиваются по мере изменения положения или ширины текущего окна.

Внутри каждого раздела данные представляются в виде двухуровневой списковой структуры (а на нижнем уровне, т.е. для гипоспиралей, в форме массива переменной длины). В итоге была реализована структура данных, схематически показанная на рис. 2.

Как показали вычисления уже с первыми вариантами программы, от 80% и более всех спиралей любой последовательности имеют длину плеча 3-4. Отсюда следует, что подавляющее большинство потенциальных макросостояний содержат по одному микросостоянию, случаи нескольких микросостояний в одном макросостоянии весьма редки. Поэтому в реализованной структуре данных основой описания является дискретное пространство всех возможных микросостояний текущего окна, которые уже затем группируются в макросостояния по признаку совпадения диаграммы.

В построенном таким образом пространстве макросостояний нет необходимости проверять их непустоту; всевозможные реализации всех макросостояний известны по построению. Процедура нахождения окрестности текущего макросостояния и моделирования перехода к одному из соседних макросостояний (либо сдвига одной из границ окна) благодаря группировке в разделы по длине диаграммы реализуется быстро, что очень важно, поскольку, как показали вычисления, число медленных переходов без изменения границ окна может достигать десятков тысяч, включая циклы непредсказуемой длины. Именно это заставило нас отказаться от очевидной альтернативной идеи — не строить заранее всё множество микро- и макросостояний

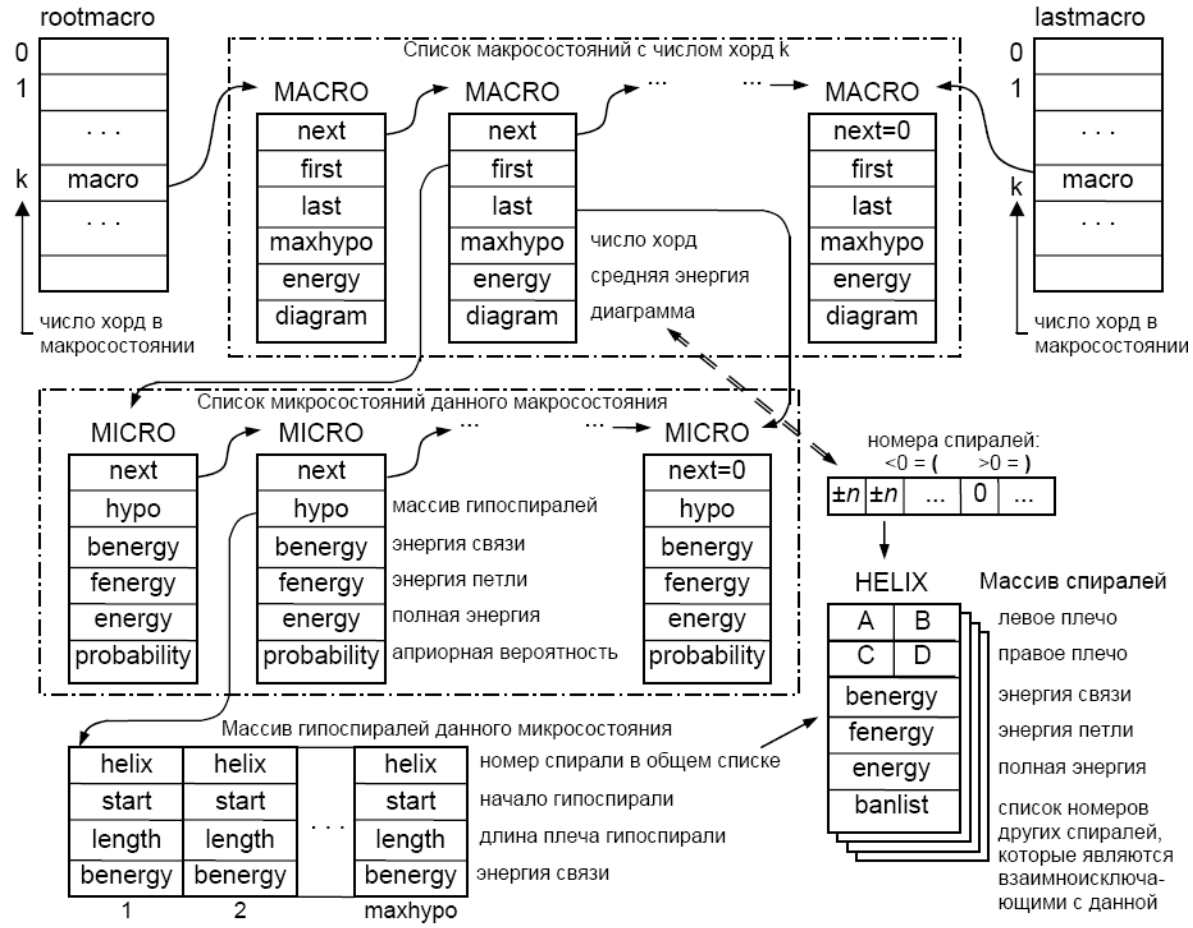

Рис. 2. Структура данных, описывающих состояние модели

текущего окна, а выполнять построение по мере необходимости, т.е. пополнять изначальное пустое множество макросостояний лишь ближайшими, отличающимися не более чем на одну хорду, соседями очередного достигнутого макросостояния, одновременно строя для каждого их них полное множество его микросостояний.

В качестве примера в таблице 2 для последовательности Sa_trpS (ген *trpS* у *S. avermitilis* MA-4680) приведены данные о распределении спиралей по длине и средние значения числа микро- и макросостояний, все в зависимости от ширины окна. Конкретно, окно с каждым из приведенных в таблице 2 значений ширины устанавливалось во всевозможных позициях на последовательности, в каждом случае находились все спирали, попадающие в это окно, и подсчитывалось число возможных микро- и макросостояний для данного окна. После этого уже для всей последовательности проведено усреднение отдельно по каждой длине плеча спирали, а также по числу состояний; полученные значения указаны в таблице.

Таблица 2. Численные характеристики потенциально возможных спиралей.

Ширина окна	Среднее число спиралей с длиной плеча:					Среднее число	Среднее число
(нуклеотидов)	3	4	5	6	>6	макросостояний	микросостояний
10	0.09	–	–	–	–	0.09	0.09
20	1.93	0.55	0.17	0.05	–	2.91	2.92
30	5.48	1.73	0.60	0.38	0.27	17.35	18.31
40	11.02	3.73	1.54	0.68	0.69	105.2	116.4
50	17.89	6.16	2.92	1.00	1.27	501	578
60	24.91	8.42	4.35	1.27	1.80	1981	2325
70	32.15	10.82	5.69	1.59	2.16	8265	9887
80	39.56	13.00	7.04	1.92	2.44	33713	40801
90	53.55	17.55	9.18	3.09	3.73	219097	284627

Для окон длиной до 40-50 нуклеотидов после сдвига рибосомы или полимеразы удобнее оказывается строить вышеописанную структуру пространства макросостояний заново. Однако при дальнейшем увеличении длины окна, с учетом наблюдаемого в таблице 2 быстрого роста всех характеристик построение структуры заново оказывается слишком трудоемким, требуя иногда нескольких часов машинного времени. Поскольку для получения вероятности терминации/антитерминации с необходимой точностью обычно требуется 10^3-10^4 прогонов алгоритма при каждом значении концентрации c (которое обычно изменялось с шагом 0,01 в интервале от 0 до 1), такой способ построения описания состояния модели сильно осложнял моделирование, особенно для сравнительно длинных последовательностей. Поэтому во *второй версии программы*, начиная с некоторой пороговой ширины окна, после сдвига правой или левой границ окна текущее пространство макросостояний не строилось заново, а наоборот, перестраивалось из достигнутого, что хотя и сложнее алгоритмически, но дает существенное увеличение производительности.

В качестве альтернативы рассматривался также *третий способ реализации алгоритма*, при котором для всей исходной последовательности заранее находилось (и заносилось в базу данных) множество всех микро- и макросостояний, а затем в процессе переходов сначала выделяется подмножество в пространстве всех макросостояний, относящееся к текущему окну, после чего реализующие его микросостояния сужаются с учетом границ окна. При массовых расчетах этот способ представляется перспективным по общим затратам времени, однако он требует с самого начала привлечения высокопроизводительных мощных СУБД и хорошо оснащенных вычислительных систем, поскольку оценки дают типичный объем такой базы данных порядка сотен гигабайт, что не всегда удобно.

4. РЕЗУЛЬТАТЫ МОДЕЛЬНОГО СЧЕТА ДЛЯ СТРЕПТОМИЦЕТОВ

Счет для каждой их четырех регуляторных областей, указанных на рисунке 1, проводился с варьированием значений параметров в указанных выше пределах. Это позволило оценить важность каждого из параметров (по характеру его влияния на результат) и увидеть, какие параметры модели предположительно общие и какие, напротив, специфичные для регуляторной области.

Забегая вперед, отметим, что наиболее существенными оказались параметры L_1, r_0, α. Влияние "размеров" рибосомы и полимеразы имеет место, но оно заметно слабее и одинаковое для всех рассмотренных организмов и генов. Это позволило выбрать для них общие значения $s_0 = 12$, $s_1 = 5$. Что касается параметра p_0, то, в принципе, его значение должно быть существенно, но, как показал счет, оно не слишком сильно влияет на характер зависимости, а

приводит, в основном, лишь к сдвигу области значений. Поэтому значение $p_0 = \frac{\pi}{14}$ может быть приемлемым компромиссом. Для контрольного счета можно принять в качестве p_0 решение уравнения (19) для h и l, отвечающих фактическим характеристикам терминатора, указанного на рис. 1. Такие значения p_0 для каждого из рассмотренных стрептомицетов приведены во второй колонке таблицы 3 (данные других колонок будут объяснены ниже).

Таблица 3. Значения параметров, специфичных для организма, где C_{max} — точка "насыщения" по величине c и $c_0^* = 1/C_{max}$.

Ген	p_0	c_{max}	c_0^*
Sv_trpE	0.150	0.50	2.000
Sa_trpS	0.130	0.65	1.538
Sa_trpE	0.126	0.65	1.538
Sc_trpE	0.116	0.20	5.000

Для каждого из организмов характерная зависимость $p(c)$ наблюдалась в своем интервале значений концентрации c. Принималось $c_0 = 1$ (т.е. по оси абсцисс откладывалась величина $\frac{c}{c_0} = c$), а концентрация c варьировалась в диапазоне от 0 до 0,75 с шагом 0,05. В первой серии экспериментов при вычислении энергии связи гипоспирали не учитывалось влияние третичной структуры, полагая $\alpha = 0$, т.е. вместо формулы (4) фактически использовалась формула (3). При этом для генов Sv_trpE и Sa_trpS были получены следующие зависимости, где по оси ординат откладывается величина $1 - p(c)$ (рис. 3-4, значения прочих параметров указаны на графиках). Для сравнения на рис. 5 воспроизведен экспериментальный график для *S. venezuela*, взятый из работы [11]; видно, что характер зависимости тот же самый.

По приведенным графикам можно сделать предварительный вывод о том, что модель является более чувствительной к изменению концентрации аминокислоты при ненулевых значениях параметра L_1, что говорит в пользу принятой гипотезы о резонансном характере силы взаимодействия шпильки с полимеразой.

Наряду с этим, для двух других генов (Sa_trpE, Sc_trpE) при тех же значениях параметров подобной зависимости не наблюдается, либо ее монотонность существенно искажена. По-видимому, это является следствием недоучета влияния третичной структуры на энергию связи спиралей в данных генах. Поэтому в следующей серии экспериментов для указанных двух генов при вычислении энергии связи спирали использовалась формула (4), где вводился в действие параметр α, значения которого выбирались в интервале от 1 до 20. Полученные для генов Sa_trpE и Sc_trpE зависимости той же величины $1 - p(c)$ от концентрации при некоторых фиксированных значениях α представлены на рис. 6, 7 (значения остальных параметров указаны на графиках).

Как видно из графиков, активизация параметра α привела к восстановлению вида характерной зависимости, по крайней мере на начальном ее участке, т.е. при значениях концентрации c от 0 до некоторой точки C_{max}, в которой выполняются условия: $p(c)$ достигает максимального значения, рибосома устойчиво находится на стоп-кодоне лидерного пептида, график $p(c)$ имеет устойчиво монотонный характер. Найденные в эксперименте значения C_{max} для всех четырех генов указаны в третьей колонке таблицы 3.

Для контроля была сделана попытка использовать аналогичные значения параметра α и для первых двух генов. Эти эксперименты наглядно продемонстрировали (рис. 8, 9), что для этих организмов учет влияния третичной структуры напротив, приводит к сильному рассогласованию получаемой зависимости с найденной в экспериментах [10] и полученной на нашей модели при $\alpha = 0$ (ср. рис. 3 и 4). Ожидаемый характер монотонного убывания величины $1 - p(c)$ с ростом концентрации сохраняется только при очень малых значениях c; диапазон

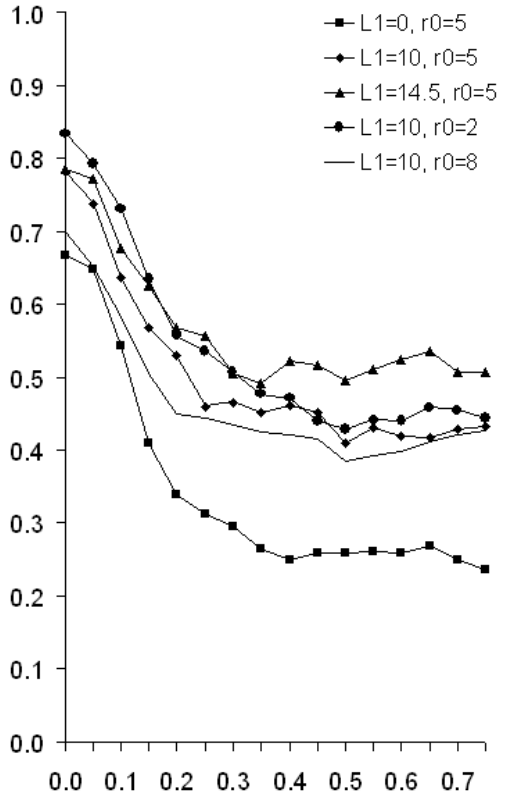

Рис. 3. Экспрессия гена Sv_trpE в зависимости от концентрации аминокислоты при различных значениях параметров модели: L_1 и r_0 (при $p_0 = 0,15$ и $\alpha = 0$)

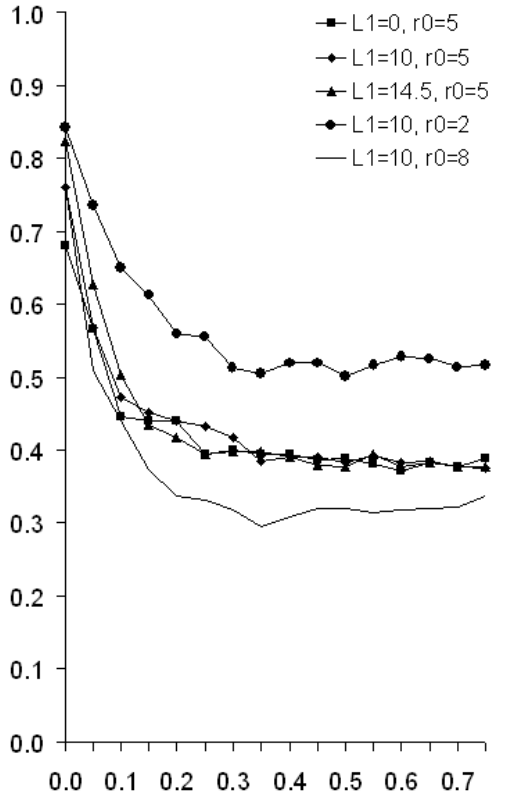

Рис. 4. Экспрессия гена Sa_trpS в зависимости от концентрации аминокислоты при различных значениях параметров модели: L_1 и r_0 (при $p_0 = 0,13$ и α=0)

изменения этой величины также значительно уменьшается. Таким образом, приходится признать, что для генов Sv_trpE и Sa_trpS влияние третичной структуры на энергию связи не должно учитываться, а для Sa_trpE и Sc_trpE его, наоборот, необходимо учесть. К сожалению, пока заранее не ясно, для каких генов необходимо вводить в модель ненулевое значение параметра α.

Несмотря на отмеченные особенности, поведение модели на различных организмах в целом выглядит достаточно согласованным. На рис. 10 показана зависимость $1 - p(c)$ одновременно для четырех рассматривавшихся генов стрептомицетов при одинаковых значениях всех параметров кроме α ($\alpha = 0$ для Sv_trpE и Sa_trpS, и $\alpha = 10$ для Sa_trpE и Sc_trpE). График каждой зависимости приведен только для интервала значений концентрации аминокислоты от 0 до C_{max} (найденные в эксперименте значения C_{max} для каждого гена указаны в таблице 3). При сравнении с экспериментальными кривыми, по-видимому, разумно рассматривать зависимость величины $1 - p(c)$ на интервале, начинающемся не в 0, а в некоторой точке $C_{min} > 0$ и сдвигать полученные графики вниз на величину репрессии.

Напомним, что по оси абсцисс графиков в действительности откладывается значение отношения $\frac{c}{c_0}$, которое численно равно c, благодаря принятой для наших экспериментов величине $c_0 = 1$. Как уже отмечалось выше, константа c_0 не имеет однозначной биологической интерпретации; более того, нет оснований полагать, что ее значение должно быть одинаковым для различных генов. Поведение модели, в частности, наличие характерного значения C_{max} концентрации аминокислоты, после которого аттенюаторный механизм регуляции вхо-

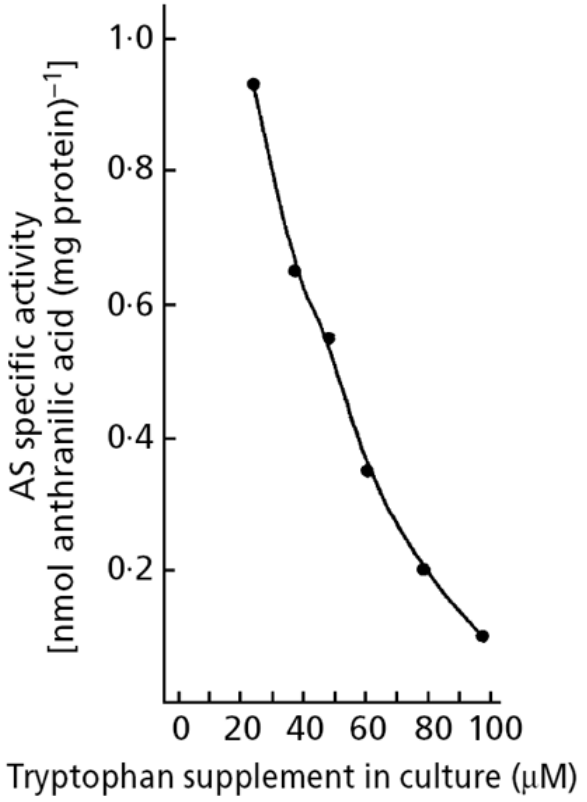

Рис. 5. График экспериментальной зависимости из [10], в которой определены 6 отмеченных точек

дит в "насыщение" и при дальнейшем увеличении концентрации, по-видимому, перестает играть определяющую роль, — позволяет предложить гипотетическую трактовку биологического смысла c_0. А именно, считать его таким значением концентрации аминокислоты в культуре, при дальнейшем увеличении которого аттенюация заметно не возрастает (точка насыщения). Предположительно, такое значение для каждого гена может быть найдено экспериментально. Аналогичные значения для нашей модели вычисляются как $c_0^* = 1/C_{max}$ и приведены в последней колонке таблицы 3. Переход к новому масштабу по оси абсцисс означает, что значение концентрации c будет меняться в интервале $[0, 1]$, одном и том же для всех генов, как изображено на рис. 11. С использованием полученных в эксперименте значений c_0 каждый такой нормированный график можно легко сопоставить по оси концентраций с экспериментальной зависимостью.

ЗАКЛЮЧЕНИЕ

В работе предложена модель аттенюаторной регуляции, которая основана на явных поддающихся анализу строго сформулированных положениях (в части описания зависимостей величин и выбора значений параметров) и по которой модельный счет на биологических примерах приводит к результатам, не расходящимся с экспериментальными данными. Предложены методики определения параметров по исходным данным. Модель реализована эффективной компьютерной программой, допускающей табличное и графическое представление результатов моделирования и широкое варьирование параметров. На основе счета получены заключения о чувствительности модели к одним параметрам и ее относительной устойчивости к другим параметрам; получены численные оценки таких биологически содержательных параметров.

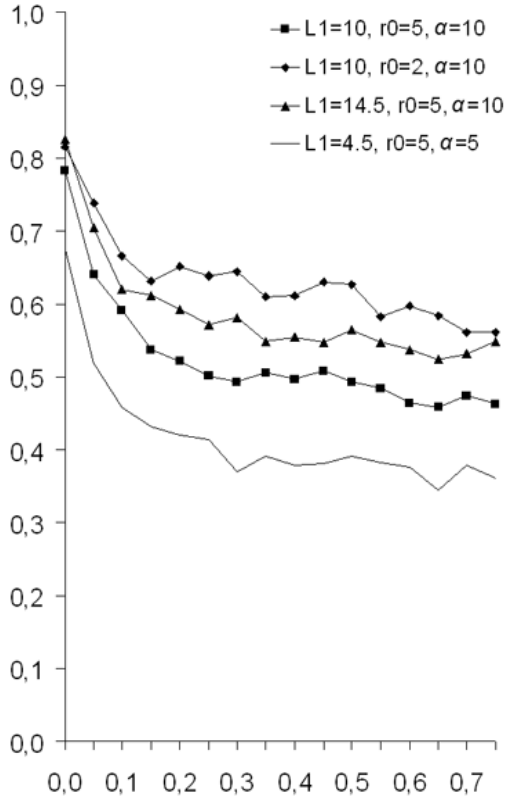

Рис. 6. Экспрессия гена Sa_trpE в зависимости от концентрации аминокислоты при различных значениях параметров модели: L_1, r_0 и α (при $p_0 = 0,126$)

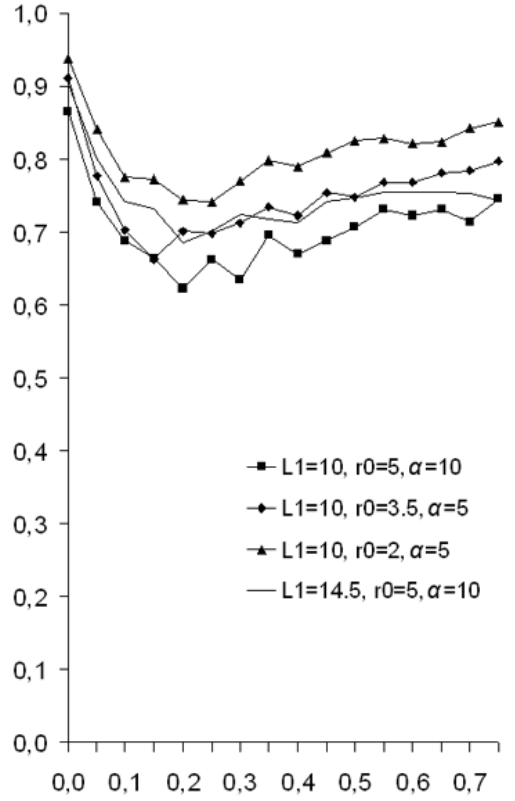

Рис. 7. Экспрессия гена Sc_trpE в зависимости от концентрации аминокислоты при различных значениях параметров модели: L_1, r_0 и α (при $p_0 = 0,116$)

Получены хотя и внутренние, но существенные для любых моделей в этой области численные характеристики: типичные длины плеч, соотношения числа микро- и макросостояний, длины циклов между двумя соседними переходами рибосомы и полимеразы и т.п. (в разделе 3).

Модель позволяет анализировать изменения (“мутации”) в исходных нуклеотидных последовательностях, сравнивать результаты таких изменений и делать из этого биологические заключения. Указанные выше графики можно представить аналитически, найти по ним характерные значения соответствующих зависимостей: точки перегиба, скорости изменения в нуле и т.п. Но, по-видимому, к этому разумно перейти после обсуждения принципов, положенных в основание модели, уточнения относительной значимости различных параметров, ее массового тестирования. Авторы предполагают представить такие результаты в другой публикации.

ПРИЛОЖЕНИЕ

Алгоритм 1. Простейший способ состоит в том, что организуется перебор всех идущих подряд троек нуклеотидов из каждой спирали носителя и для каждого такого набора троек проверяется его совместность. Если хотя бы один такой набор совместен, то он произвольным образом расширяется до (непродолжаемого) набора, т.е. микросостояния. Это требует времени экспоненциального от числа спиралей в носителе, но полиномиального от суммарной длины всех этих спиралей.

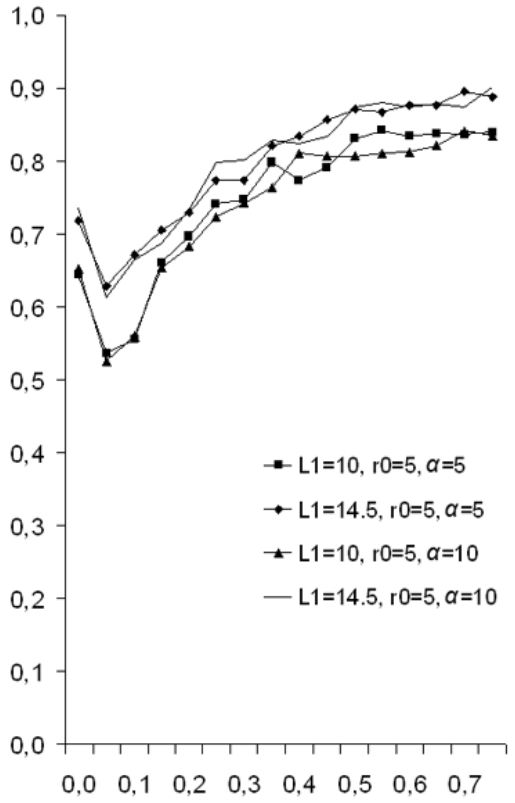

Рис. 8. Экспрессия гена Sv_trpE в зависимости от концентрации аминокислоты при различных значениях параметров модели: L_1, r_0 и α (при $p_0 = 0,15$)

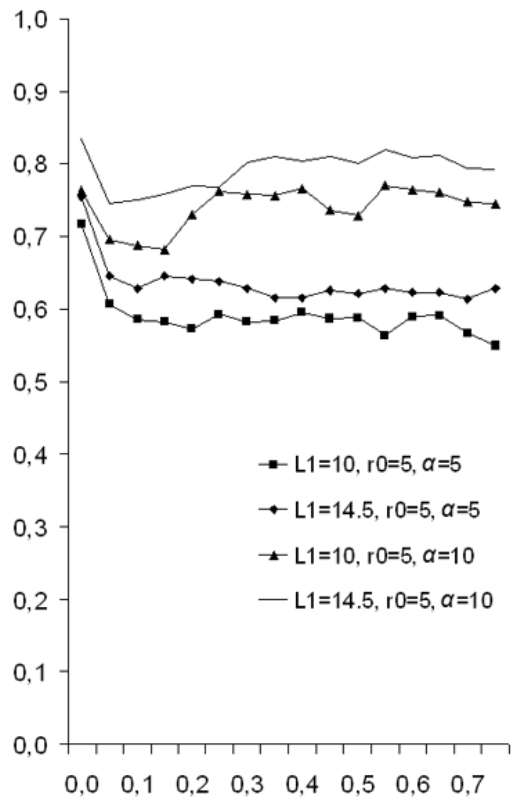

Рис. 9. Экспрессия гена Sa_trpS в зависимости от концентрации аминокислоты при различных значениях параметров модели: L_1, r_0 и α (при $p_0 = 0,13$)

Более содержательный способ (правда, тоже экспоненциальный от числа спиралей) состоит в следующем. Будем называть *областью* любую сплошную подпоследовательность исходной последовательности. Методом динамического программирования составляем список троек ⟨область R, множество M спиралей, знаки 0 или 1⟩ таких, что область R содержит хотя бы одно микросостояние, носитель которого совпадает с M и которое является *неразложимым* (т.е. его диаграмма содержит внешнюю скобку) — тогда в конце указывается знак 0, или разложимым (т.е. его диаграмма не содержит внешней скобки) — тогда в конце указывается знак 1. Пусть мы хотим узнать, содержит ли данная область R такое микросостояние. Если оно неразложимое, то сведение этого вопроса к соответствующим вопросам про меньшие области простое: если содержит, то это же микросостояние содержит и одну из двух обрезанных с краю на один нуклеотид областей, или существует внешняя спираль и три её склейки на краях данной области, а обрезанная с каждого края на три нуклеотида область содержит микросостояние либо с тем же носителем, либо с уменьшенным на внешнюю спираль (алгоритм должен перебрать всех кандидатов на внешнюю спираль из M).

Если микросостояние разложимое, то пусть мощность его носителя N равна n. Построим ориентированный граф с разметкой вершин, в котором вершинами являются пары ⟨подобласть области R, число i⟩ такие, что эта подобласть содержит некоторое неразложимое микросостояние, носитель которого является подмножеством N и имеет мощность ровно i (достаточно рассматривать минимальные по включению подобласти для каждого i). Каждая вершина помечена числом i, указанным в её второй компоненте. Проводим ребро из одной вершины в

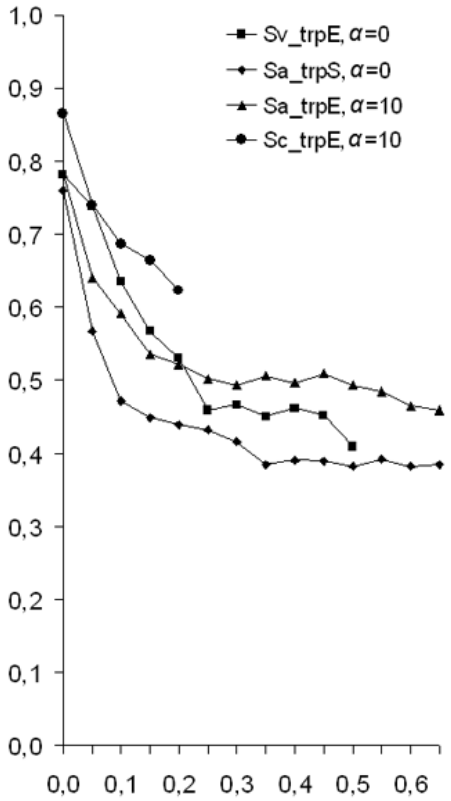

Рис. 10. Сравнение модельных зависимостей для четырех генов стрептомицетов при одинаковых значениях всех параметров ($L_1 = 10$, $r_0 = 5$ и $c_0 = 1$) кроме α

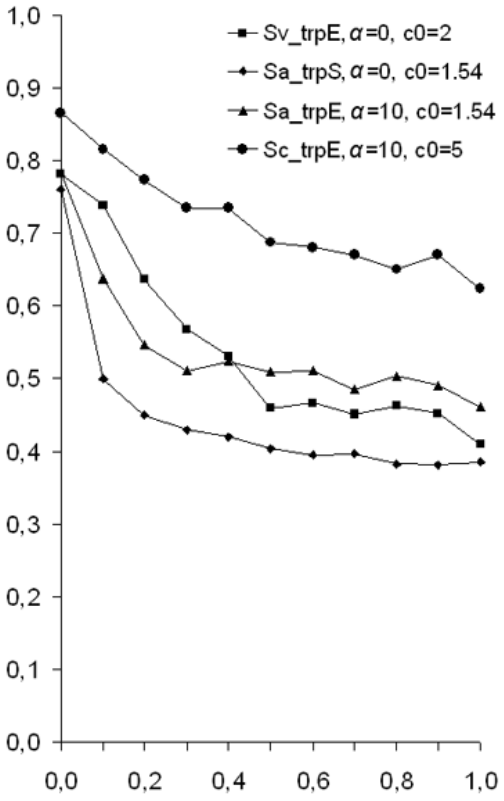

Рис. 11. Сравнение нормированных зависимостей для четырех генов стрептомицетов при одинаковых значениях всех параметров ($L_1 = 10$ и $r_0 = 5$) кроме α и c_0

другую, если первый подотрезок лежит левее второго и не пересекается с ним. Поскольку носители неразложимых компонент микросостояния не могут пересекаться, задача свелась к проверке того, имеется ли в графе путь с максимально возможной суммой пометок вершин (равной n). Наш граф ациклический, а как известно, для ациклического графа задача поиска пути с максимальной суммой пометок вершин проста (она решается динамическим программированием, где начала возможного пути перебираются так, что на каждом шагу выбирается вершина, из которых не исходит ребер в ещё не просмотренные вершины). Дойдя до всей области и данного носителя, мы ответим на исходный вопрос. Конец описания алгоритма 1.

Алгоритм 2. Очевидно, можно считать, что данная диаграмма не содержит двух пар скобок одной и той же спирали, расположенных с обеих сторон вплотную друг к другу. *Поддиаграммой* диаграммы назовём связную её часть, которая сама является диаграммой. Диаграммы бывают *разложимыми*, когда внешняя скобка отсутствует, и *неразложимыми* (в ином случае). Методом динамического программирования составляем список таких пар ⟨область исходной нуклеотидной последовательности, поддиаграмма исходной диаграммы⟩, что эта область содержит хотя бы одну реализацию этой поддиаграммы (в этом случае будем говорить, что *область реализует поддиаграмму*). Пусть мы хотим узнать, реализует ли данная область R данную диаграмму D. Если D неразложимая, то сведение этого вопроса к соответствующим вопросам про меньшие области тривиально: если реализует, то ту же диаграмму D реализует и одна из двух обрезанных с краю на один нуклеотид областей или обрезанная с каждого края на три нуклеотида область реализует диаграмму, получающуюся из D удалением пары

внешних скобок (причём на этих шести нуклеотидах действительно должны быть три склейки внешней спирали).

Если же D разложима, то пусть она разлагается на n неразложимых диаграмм. Построим n-дольный ориентированный граф, в котором вершинами i-ой доли являются подобласти области R, реализующие i-ую (слева) диаграмму (достаточно рассматривать минимальные по включению подобласти для каждой из этих n диаграмм). Проводим ребро из вершины i-ой доли в вершину $(i+1)$-ой доли, если первая подобласть лежит левее второй и не пересекается с ней. Теперь задача свелась к тривиальной: проверить, есть ли в графе хотя бы один путь из 1-ой доли в n-ую.

Если мы хотим не просто проверять непустоту диаграммы, а ещё и составлять список всех её микросостояний, то нужно добавить к паре ⟨область, поддиаграмма⟩ ещё и список реализующих эту поддиаграмму микросостояний. Для того, чтобы восстанавливать этот список, следует в указанном графе перебирать все пути из 1-ой доли в n-ую. Чтобы в нашем списке микросостояний не было повторов, важно, во-первых, при объединении множеств микросостояний из двух обрезанных областей (для неразложимой диаграммы) выбрасывать повторы (либо вместо этих двух брать все минимальные по включению подобласти, реализующие данную диаграмму), а во-вторых, для разложимой диаграммы использовать в графе лишь минимальные по включению подобласти, допускающие соответствующую диаграмму.

Алгоритм 3. Вычисление сумм в формуле (8) при *уменьшении макросостояния* выполняется следующим образом. Пусть S — спираль, из которой удаляется некоторая гипоспираль. Перебираем микросостояния ω из Ω и для каждого из них смотрим, какова суммарная энергия склеек соответствующей гипоспирали из S. Попутно считаем сумму произведений $p(\omega)$ на вероятность расклейки этой гипоспирали (эта вероятность зависит только от упомянутой энергии). Эта сумма равна искомой. Заметим, что при этом не требуется перебирать микросостояния из Ω'.

Вычисление этих сумм при *увеличении* диаграммы выполняется следующим образом. Пусть S — спираль, из которой будет добавлена одна гипоспираль (ее находим по микросостоянию ω'). Перебираем микросостояния ω из Ω и для каждого из них смотрим, какова вероятность появления гипоспирали из S на соответствующей паре участков (поскольку максимально продолженная гипоспираль определяется однозначно, эту вероятность можно считать по формуле, приведённой в тексте или в стиле [1]-[4]) как сумму вероятностей по всем парам нуклеотидов, которые могут дать начало новой гипоспирали, а для каждой пары вероятность их встречи зависит лишь от того, насколько они прижаты друг к другу имеющимися склейками, что и отражено в упомянутой формуле). Попутно считаем сумму произведений $p(\omega)$ на упомянутые вероятности. Эта сумма равна искомой, и при этом не требуется перебирать микросостояния из Ω'.

Списки всех микросостояний, реализующих любое появившееся макросостояние лучше составить до подсчета этих сумм (это можно сделать заранее для всех макросостояний, используя распараллеливание для всех них). Если полный перебор микросостояний затруднителен, можно воспользоваться следующим процессом стохастического моделирования. Для каждого реализуемого макросостояния находим одно из его микросостояний и начинаем случайное блуждание по микросостояниям, пользуясь вышеприведёнными формулами для вероятностей перехода между микросостояниями. Соседним считаем микросостояние, отличное от данного и получающееся из него одной или двумя крайними расклейками в разных гипоспиралях с последующим продолжением гипоспиралей, оказавшихся из-за этих расклеек не продолженными до конца (достаточность двух расклеек следует из предложения 1). Для ускорения процесса можно накладывать дополнительные требования на множество соседних микросостояний, например, упорядоченным образом менять вершины дерева, к которым примыкают расклейки,

запрещать многократные повторы, и т.д. Набрав достаточный статистический материал, составляем список микросостояний, их приближённых вероятностей, вероятностей расклеек или склеек тех или иных гипоспиралей и других необходимых нам сведений.

Поскольку отрезок исходной последовательности с интересующими нас микросостояниями будет меняться то предпочтительно для каждого макросостояния рассматривать "канонический" для него отрезок — от начала левого плеча спирали, соответствующей самой левой скобке до конца правого плеча спирали, соответствующей самой правой скобке. При переборе микросостояний на этом отрезке одновременно будут перебираться и микросостояния, "умещающиеся" на его подотрезках.

Благодарности

Авторы глубоко благодарны А.А. Миронову за многочисленные разъяснения по теме работы и за постановку рассмотренной здесь задачи перед одним из авторов. Авторы глубоко благодарны М.С. Гельфанду, на семинаре которого была доложена эта работа, за интерес и ценные критические замечания.

СПИСОК ЛИТЕРАТУРЫ

1. Кистер А.Э., Миронов А.А., Дроздов-Тихомиров Л.В. Количественная кинетическая модель гидролиза-синтеза АТР мембранной H-ATP-АЗОЙ. *Молекулярная биология*, 1984, том 18, №6, стр. 1476-1485.
2. Миронов А.А., Кистер А.Э. Теоретический анализ кинетики образования вторичной структуры РНК в процессе транскрипции и трансляции. Учет дефектных спиралей. *Молекулярная биология*, 1985, том 19, №5, стр. 1350-1357.
3. Миронов А.А., Кистер А.Э. Теоретический анализ структурных перестроек в процессе образования вторичных структур РНК. *Молекулярная биология*, 1989, том 23, №1, стр. 61-71.
4. Mironov A.A., Lebedev V.F. A kinetic model of RNA folding. *BioSystems*, 1993, vol. 30, pp. 49-56.
5. Сингер М., Берг П. *Гены и геномы.* М.: Мир, 1998. (Singer M., Berg P. *Genes and genomes.* United States, University Science Book, 1990.)
6. Gorbunov K.Yu., Lyubetsky V.A. A model of tryptophan biosynthesis regulation. *Proceedings of the fourth International Conference on Bionnformatics of Genome Regulation and Structure, BGRS'2004*, Novosibirsk, Russia, July 25-30, 2004, vol. 2, pp. 53-55.
7. Yin H., Artsimovitch I., Landick R., Gelles J. Nonequilibrium mechanism of translation termination from observations of single RNA polymerase molecules. *PNAS*, 1999, vol. 96, no. 23, pp. 13124-13129.
8. Wilson K., and von Hippel P. Transcription termination at intrinsic terminators: the role of the RNA hairpin. *Proc. Natl. Acad. Sci. U. S. A.*, 1995, vol. 92, pp. 8793-8797.
9. Lynn, S., Kasper, L., and Gardner, J. Contributions of RNA secondary structure and length of the thymidine tract to transcription termination at the thr operon attenuator. *J. Biol. Chem.*, 1988, vol. 263, pp. 472-479.
10. Cong Lin, Ashish S. Paradkar and Leo C. Vining. Regulation of an anthranilate synthase gene in Streptomyces venezuelae by a trp attenuator. *Microbiology*, 1998, vol. 144, pp. 1971-1980.
11. Любецкий В.А., Селиверстов А.В. Регуляция экспрессии генов биосинтеза аминокислот и аминоацил-тРНК синтетаз у актинобактерий. *Молекулярная биология*, 2005 (в печати).