БИОФИЗИКА, 2016, том 61, вып. 1, с. 40–49

МОЛЕКУЛЯРНАЯ БИОФИЗИКА

УДК 577.3

СТОХАСТИЧЕСКАЯ МОДЕЛЬ ФОРМИРОВАНИЯ МОЛЕКУЛЯРНЫХ КОНФИГУРАЦИЙ ЭНХАНСЕРА

© 2016 г. Г.М. Демидов*, М.Г. Самсонова, В.В. Гурский** *****

**Санкт-Петербургский национальный исследовательский академический университет РАН, 194021, Санкт-Петербург, ул. Хлопина, 8/3;*

***Санкт-Петербургский политехнический университет Петра Великого, 195251, Санкт-Петербург, Политехническая ул., 29;*

****Физико-технический институт им. А.Ф. Иоффе, 194021, Санкт-Петербург, Политехническая ул., 26*

E-mail: gursky@math.ioffe.ru

Поступила в редакцию 13.10.15 г.

Комбинации свободных и занятых сайтов присоединения транскрипционных факторов в энхансере, регулирующем транскрипционную активность гена-мишени, формируют различные молекулярные конфигурации. В рамках термодинамических моделей экспрессии гена вероятность активации транскрипции гена-мишени вычисляется через вероятности молекулярных конфигураций энхансера. В работе представлена простая стохастическая модель формирования таких конфигураций, учитывающая процессы неспецифичного связывания транскрипционных факторов с ДНК, скольжение белков вдоль ДНК и специфичное присоединение к энергетически выгодному сайту. С помощью модели исследованы свойства молекулярных конфигураций регуляторного района гена *knirps*, экспрессирующегося в ходе раннего развития эмбриона дрозофилы, при регуляции транскрипционными факторами Hunchback и Bicoid. В рамках принятых в модели допущений показано, что процесс присоединения транскрипционных факторов в регуляторном районе может быть разбит на три последовательные динамические стадии с разной скоростью присоединения. Сайты связывания транскрипционных факторов можно разделить на три группы по степени занятости. Рассчитанная вероятность занятости сайтов в состоянии квазиравновесия значительно отличается от распределения, возникающего в рамках предположения о термодинамическом равновесии и использующегося в термодинамических моделях экспрессии. Полученные результаты могут служить первым шагом для уточнения моделей экспрессии с помощью более детальной информации о регуляторных районах генов-мишеней.

Ключевые слова: энхансер, стохастическое моделирование, термодинамическая модель, дрозофила, гены сегментации, knirps, Hunchback, Bicoid.

Важным направлением в моделировании транскрипционной регуляции экспрессии генов эукариот являются термодинамические модели, основанные на применении методов статистической термодинамики [1–8]. В этих моделях вероятность активации гена-мишени вычисляется в зависимости от различных состояний регуляторного района (энхансера), в которых учитывается, какие сайты связывания транскрипционных факторов (ССТФ) заняты своим белком, а какие свободны. Вероятность связывания транскрипционного фактора (ТФ) с сайтом определяется через статистический вес этого сайта, который зависит от энергии присоединения для данного ТФ. Состояния энхансера называются молекулярными конфигурациями и определяются как комбинации свободных и занятых ССТФ в данном состоянии. Совокупность статистических весов ССТФ, занятых в данной конфигурации, определяет вероятность такой конфигурации. В предположении термодинамического равновесия для системы взаимодействующих белков и ДНК вероятности молекулярных конфигураций вычисляются с использованием распределения Больцмана.

Реальное распределение для вероятностей конфигураций энхансера может отличаться от распределения, возникающего из термодинамических моделей, по двум причинам. Во-первых, динамика процессов формирования конфигураций может приводить к тому, что промежуточные (нестационарные) конфигурации будут связаны с привлечением полимеразы на базальный промотор и соответствующим началом транскрипции. Это означает, что такие промежуточные состояния

Сокращения: ССТФ – сайты связывания транскрипционных факторов, ТФ – транскрипционные факторы, PWM – матрица позиционных весов, Hb – Hunchback, Bcd – Bicoid.

должны учитываться при моделировании экспрессии. Во-вторых, в распределении Больцмана учитываются только энергии связывания для ССТФ и не учитывается история формирования конфигураций, определяемая процессами поиска ТФ своих сайтов специфичного связывания. Эти процессы включают в себя диффузию молекул ТФ в нуклеоплазме, неспецифичное связывание с хроматином, одномерное скольжение вдоль ДНК и специфичное связывание с ССТФ [9]. Учет только энергий связывания ССТФ (аффинностей сайтов) при расчете квазистационарного распределения состояний энхансеров соответствует предположению о том, что главным лимитирующим процессом при формировании этих состояний является диффузия. Однако экспериментальные исследования на эукариотических клетках показывают, что таким лимитирующим фактором является процесс неспецифичного связывания [10–12]. Таким образом, явный учет процессов поиска транскрипционным фактором специфичных сайтов может приводить к конфигурациям регуляторных районов, в которых вероятности занятости отдельных сайтов будут в большей степени зависеть от различных геометрических и динамических эффектов и в меньшей степени от аффинности сайта, определяемой исключительно его нуклеотидным составом.

В работе предлагается метод расчета вероятностей конфигураций энхансера на основе стохастической модели формирования конфигураций. В модели рассмотрены все основные процессы, ведущие к событиям связывания или освобождения ССТФ в регуляторной последовательности. Результат 3D-диффузии молекул ТФ в нуклеоплазме выражается в виде равномерного распределения для вероятностей достижения молекулами произвольной позиции в ДНК. Модель учитывает стохастические процессы неспецифичного связывания ТФ с ДНК и отсоединения от ДНК, а также 1D-транспорт молекулы ТФ вдоль ДНК.

Модель апробируется на простой системе, состоящей из регуляторного района одного гена и двух транскрипционных факторов. В качестве гена-мишени выбран ген *knirps* (*kni*), который является одним из генов сегментации, контролирующих формирование пространственного паттерна у плодовой мушки дрозофилы вдоль главной, антериорно-постериорной оси эмбриона ([13,14]). Ген *kni* экспрессируется в эмбрионе в виде антериорного и абдоминального доменов [14]. В качестве регуляторов в модели рассмотрены ТФ Hunchback (Hb), являющийся репрессором гена *kni*, и Bicoid (Bcd), активирующий *kni*. Выбранные ген и транскрипционные факторы являются частью регуляторной сети генов семейства gap, термодинамическая модель которой была представлена ранее [8].

В результате стохастического моделирования были проанализированы основные закономерности физического процесса формирования молекулярных конфигураций регуляторного района гена *kni*. Показано, что динамика изменения конфигураций имеет разную скорость при разном числе занятых сайтов. Вычисления привели к распределению вероятности занятости сайта, при котором большинство сайтов с примерно равной вероятностью являются либо связанными с ТФ, либо свободными. Значительная часть сайтов остается почти всегда свободной, и еще одна часть сайтов имеет высокую вероятность быть занятой. Полученное распределение значительно расходится с распределением, ожидаемым из термодинамических моделей. Результаты позволяют сделать вывод о необходимости модификации принятого в этих моделях формализма для описания процессов взаимодействия ТФ и регуляторных районов ДНК.

МАТЕРИАЛЫ И МЕТОДЫ

Регуляторная последовательность и сайты присоединения транскрипционных факторов. Модель симулирует стохастическую динамику присоединения ТФ к регуляторному району (энхансеру) гена *kni* у плодовой мушки *Drosophila melanogaster*. Процедура выбора регуляторного района и ССТФ полностью совпадает с таковой из работы [8]. В качестве потенциального регуляторного района была выбрана окрестность гена *kni* в референсном геноме, отсчитывающая 12 тыс. п.н. от сайта инициации транскрипции в направлении 5′. В этой последовательности были выделены участки, доступные для действия ДНКазы I. Такие участки, предположительно, являются открытыми для присоединения ТФ [15]. В качестве транскрипционных факторов, взаимодействующих с рассмотренным регуляторным районом, были выбраны белки Bcd и Hb. Оба транскрипционных фактора регулируют транскрипционную активность гена *kni* в ходе раннего развития дрозофилы, при этом Bcd является активатором, а Hb – репрессором [8,14,16]. Сайты специфичного связывания для этих ТФ в регуляторной последовательности находились с использованием матриц позиционных весов (PWM; http://autosome.ru/iDMMPMM), при этом энергия связывания ТФ для данного сайта вычислялась как $\exp(w)$, где w есть значение веса (PWM-score) для этого сайта, который вычисляется с помощью PWM и отсчитывается от веса сильнейшего сайта [17]. Специфичные ССТФ характеризуются высоки-

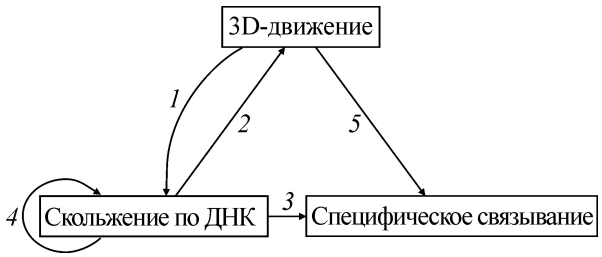

Рис. 1. Процессы в стохастической модели формирования конфигураций регуляторного района. Белки могут участвовать в 3D-движении (вне ДНК), в 1D-движении (скольжение по ДНК) и быть специфично связанными со своими ССТФ. Стрелками показаны следующие реакции, переключающие белки между этими процессами и состояниями: неспецифичное связывание белка с регуляторным районом в результате присоединения к ДНК из нуклеоплазмы (реакция *1*), отсоединение белка от регуляторного района (*2*), специфичное связывание белка в результате скольжения по ДНК (*3*), скольжение белка по ДНК (4), специфичное связывание белка с регуляторным районом в результате присоединения к ДНК из нуклеоплазмы (*5*).

ми значениями энергии связывания. Всего на обеих нитях ДНК в регуляторном районе было найдено 135 специфичных сайтов для двух ТФ, из них 54 сайта для Bcd и 81 сайт для Hb.

Алгоритм стохастической симуляции. Стохастическая симуляция процесса формирования конфигураций регуляторного района гена *kni* происходит по модифицированному методу Гиллеспи [18]. Модель включает в себя следующие процессы (рис. 1): 3D-движение молекул ТФ, неспецифичное присоединение ТФ к регуляторному району ДНК, 1D-скольжение неспецифично связанных молекул ТФ вдоль ДНК, отсоединение молекул ТФ от ДНК, специфичное связывание молекул ТФ с их сайтами. 3D-движение ТФ в нуклеоплазме явно не моделируется; вместо этого предполагается, что в результате диффузии молекулы белков оказываются в окрестности ДНК с одинаковой вероятностью для любой позиции в регуляторном районе.

Общее число молекул ТФ в системе постоянно и равно $N = 1000$ (по 500 молекул каждого из двух ТФ). В каждый момент времени t выполняется: $N = N_A(t) + N_B(t)$, где $N_A(t)$ – число молекул вне ДНК, $N_B(t)$ – число молекул на ДНК; в начальный момент времени $t = 0$: $N_A(0) = N$, $N_B(0) = 0$. Время $t + \Delta t$ каждого следующего события вычисляется из экспоненциального распределения для интервалов Δt с плотностью вероятности $f(t,\tau^{-1}) = \tau^{-1}e^{-t/\tau}$, где, для упрощения сравнения с экспериментальными данными, характеристическое время τ выбирается по-разному, в зависимости от того, происходит ли событие с молекулами вне ДНК или с молекулами на ДНК: $\tau = \tau_A = a/N_A(t)$ для реакций связывания (реакции *1* и *5* на рис. 1) и $\tau = \tau_B = b/N_B(t)$ для реакций на ДНК (реакции *2*–*4* на рис. 1). Значения параметров a и b выбраны эмпирически: $a = 10^6$, $b = 0{,}005$. После каждого процесса присоединения ТФ к ДНК и отсоединения от ДНК происходит пересчет соответствующих количеств молекул $N_A(t)$ и $N_B(t)$ и характеристических времен τ_A и τ_B. Рассчитанные времена событий записываются по ходу симуляции в общую очередь с сохранением информации о том, к какому из двух множеств белков (связанным или не связанным с ДНК) относятся эти события.

Стохастическая симуляция происходит согласно следующему алгоритму.

1. В начальный момент времени случайным образом выбирается сайт S в регуляторном районе, в котором происходит связывание одной молекулы ТФ с ДНК. Если этот сайт является сайтом специфичного связывания, молекула остается там до конца симуляции (реакция *5* на рис. 1). В противном случае связывание характеризуется как неспецифичное (реакция *1* на рис. 1).

2. После неспецифичного связывания происходит процесс 1D-скольжения этой молекулы вдоль ДНК (реакция *4* на рис. 1), для которого выбирается направление движения и длина шага в ходе одного акта скольжения. При выборе направления предполагается, что скольжение происходит с большей вероятностью в сторону той локальной окрестности на последовательности, которая более выгодна с точки зрения энергии связывания с этим белком. Чтобы найти такую окрестность, вводится понятие левого и правого потенциалов связывания (φ_l и φ_r соответственно) относительно сайта неспецифичного связывания S, представляющих усредненные значения весов нескольких отрезков последовательности слева и справа от неспецифичного сайта S:

$$\varphi_l(S) = \frac{1}{L}\sum_{j=i-L}^{i-1} w(j), \qquad \varphi_r(S) = \frac{1}{L}\sum_{j=i+1}^{i+L} w(j),$$

где $w(j)$ – PWM-вес потенциального ССТФ, начинающегося с позиции j в ДНК (порядковый номер нуклеотида в регуляторном районе) и имеющего длину, равную длине мотива рассматриваемого ТФ; через i обозначена начальная позиция сайта S, и L есть глубина сканирования окрестности сайта S для вычисления

потенциалов связывания (L = 5 нуклеотидов). Направление скольжение выбирается следующим образом: генерируется случайное число $0 \le x \le 1$ из равномерного распределения; молекула движется вдоль ДНК налево от сайта S, если относительная энергия присоединения левой окрестности сайта S больше порога x, т.е. если выполняется неравенство:

$$x < \frac{\exp(c\varphi_l(S))}{\exp(c\varphi_l(S)) + \exp(c\varphi_r(S))},$$

где c – эмпирический параметр (c = 0,1). Если это неравенство не выполняется, белок движется по ДНК направо. Использование случайного порога x вместо фиксированного позволяет осуществлять движение в энергетически невыгодную сторону. Длина скольжения λ вычисляется как линейная функция от потенциала: $\lambda = \alpha\varphi + \beta$, где $\varphi = \varphi_l$. если движение происходит налево, и $\varphi = \varphi_r$, если направо. Значения параметров (α = –0,75, β = 1) выбирались так, чтобы разброс значений потенциалов, который получался в ходе вычислений для многих сайтов неспецифичного связывания, соответствовал разбросу длины скольжения от минимально возможной длины в один нуклеотид до максимально разрешенной длины в десять нуклеотидов. Если в ходе скольжения на пути белка встречается другая молекула, связанная с ДНК, скольжение прекращается, белок останавливается рядом с уже присоединенной молекулой и происходит переход к следующему шагу алгоритма.

3. Если после совершения белком одномерного движения вдоль ДНК он оказывается на сайте специфичного связывания, белок остается там до конца симуляции (реакция *3* на рис. 1). В противном случае сразу после скольжения производится тест на отсоединение от ДНК, в результате которого молекула может вернуться в режим 3D-движения (реакция *2* на рис. 1). Отсоединение может произойти, если белок оказывается на новом неспецифичном сайте связывания с низким весом (энергетически невыгодный сайт). Для более непрерывной оценки влияния веса на вероятность отсоединения вводится сигмоидальная функция $t(w(j)) = 1/(1 + e^{-\theta(w(j) - \gamma)})$, где $w(j)$ – PWM-вес сайта, в который переместился белок после скольжения и который начинается с нуклеотида с номером j, и θ = –0,3, γ = –0,15 – параметры, эмпирически подобранные исходя из распределения значений $w(j)$. После расчета $t(w(i))$ генерируется случайное число $0 \le x \le 1$ из равномерного распределения, и отсоединение белка от ДНК происходит в случае $x > t(w(i))$. В противном случае белок остается присоединенным к ДНК. Использование случайного порога x вместо фиксированного позволяет белкам оставаться связанными с ДНК в энергетически невыгодных сайтах.

4. Далее вычисляются времена $t + \Delta t$ новых событий для белков, находящихся в 3D-движении (с характеристическим временем τ_A), и для белков, находящихся на ДНК (с характеристическим временем τ_B), и добавляются в общую очередь. После этого проверяется запись из очереди с ближайшим значением времени и выясняется, к каким из двух множеств белков (связанным или не связанным с ДНК) относится событие, соответствующее этому времени. Если это событие связано с молекулами, находящимися в 3D-движении, то производится тест на очередное присоединение одной молекулы к ДНК, и далее повторяется весь алгоритм (шаги 1–4). Если событие связано с молекулами на ДНК, то случайным образом выбирается одна из таких молекул и с ней повторяются шаги 2–4 (скольжение по ДНК, тест на специфичное связывание, тест на отсоединение, генерация нового времени).

Механизм репрессии. Для белка Hb реализуется механизм короткодействующей репрессии. Он заключается в том, что, связываясь специфично с ДНК, молекула белка-репрессора способна деформировать локальную окрестность своего сайта связывания, не позволяя в этой окрестности присоединяться любым другим молекулам. Если в запрещенную окрестность попадают сайты присоединения белков-активаторов, происходит эффективное ослабление их общей активирующей способности. В модели реализуются два способа специфичного связывания для белка Hb [4]. Если в результате реакций *3* или *5* на рис. 1 молекула Hb связывается со своим сайтом, слева и справа от которого в пределах нуклеотидов нет занятых ССТФ, присоединение репрессора считается «активным», так что в соответствующей окрестности запрещается любое связывание белков и их скольжение. Если в этой окрестности в момент специфичного связывания белка Hb уже находятся другие связанные белки, специфичное связывание Hb не сопровождается модификацией этой окрестности, т.е. присоединение Hb считается «неактивным» (репрессии не происходит). Для эффективного радиуса репрессии δ принято значение δ = 100 п.н., близкое к значениям, использованным ранее в модели сети генов gap [8], и вытекающее из результатов экспериментов [19].

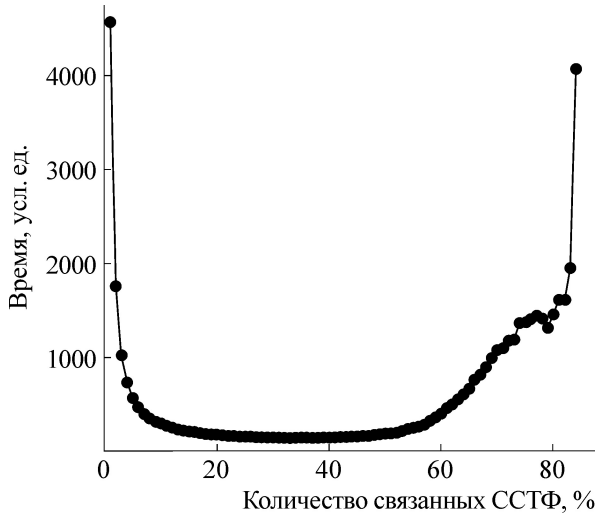

Рис. 2. Динамика формирования конфигурации регуляторного района. Показано время (в условных единицах) между двумя последовательными событиями специфичного связывания ТФ в регуляторном районе как функция от количества занятых специфичных ССТФ (в процентах от общего числа сайтов).

Критерий остановки симуляции. Симуляция завершается при достижении системы состояния равновесия, а именно при выполнении двух условий: число занятых сайтов специфичного связывания составляет более чем 50% от общего их количества, и время Δt, прошедшее с момента последнего события специфичного связывания, превышает определенный порог. В качестве такого порога было выбрано значение 15000 единиц, соответствующее примерно 15 с в клетке. Это время близко к экспериментально измеренному характеристическому времени специфичного связывания для ТФ в эмбриональных стволовых клетках [11].

Значения параметров в модели. Приведенные выше значения свободных параметров в стохастическом моделировании подбирались так, чтобы удовлетворить следующим экспериментальным наблюдениям. Во-первых, соотношение свободных (совершающих 3D-движение) и связанных с ДНК белков должно составлять примерно 80 на 20% [12]. Далее, среднее время нахождения белков в состоянии 3D-движения составляет примерно 4 с, среднее время нахождения на ДНК примерно 1 с и среднее время специфичного связывания примерно 15 с (последнее время использовалось только для оценки временно́го порога в критерии остановки, как описано выше) [11]. 1000 единиц времени в модели соответствовали примерно 1 с реального времени в клетке.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Динамика формирования конфигурации. Результаты стохастического моделирования показывают, что динамика формирования конфигурации регуляторного района *kni* имеет немонотонный характер и состоит из трех характерно выраженных, последовательных стадий (рис. 2). На первой стадии следующие друг за другом события связывания новых специфичных ССТФ разделены значительными временны́ми интервалами. Это объясняется тем, что неспецифично связанных белков на ДНК еще мало, и они проводят там относительно короткое время (в среднем около 1 с), за которое они не успевают обнаружить сайт специфичного связывания. Такой эффект согласуется с экспериментальными оценками, согласно которым число попыток найти специфичный сайт, сопровождающихся последовательными процессами неспецифичного связывания и отсоединения от ДНК, может быть порядка 100 [11]. Существование такой стадии в динамике также можно интерпретировать как проявление того факта, что лимитирующим шагом в поиске специфичных ССТФ является неспецифичное связывание ТФ с ДНК [10–12]. В ходе первой стадии происходит связывание примерно 10% всех специфичных ССТФ.

По мере увеличения числа событий неспецифичного связывания в динамике проявляется следующая стадия, в которой новые специфичные сайты занимаются почти с постоянной скоростью, при этом время между последовательными событиями связывания специфичных сайтов имеет минимальное значение. В ходе этого периода занимается наибольшая часть ССТФ от начальных 10% до примерно 60% сайтов.

После достижения примерно 60% занятых специфичных ССТФ начинается третья стадия в формировании конфигурации энхансера, которая характеризуется постепенным увеличением временно́го интервала между последовательными событиями связывания. Такая динамика является следствием общего насыщения регуляторного района молекулами ТФ, так что новым молекулам все сложнее найти свободный специфичный сайт. Переходный период до полного установления конфигурации длится примерно до достижения 80% занятых сайтов, после чего время ожидания следующего события связывания специфичного сайта резко возрастает (рис. 2). В конце третьей стадии можно считать, что регуляторный район достиг квазистационарного состояния.

Биофизические характеристики репрессии и скольжения по ДНК. С помощью стохастиче-

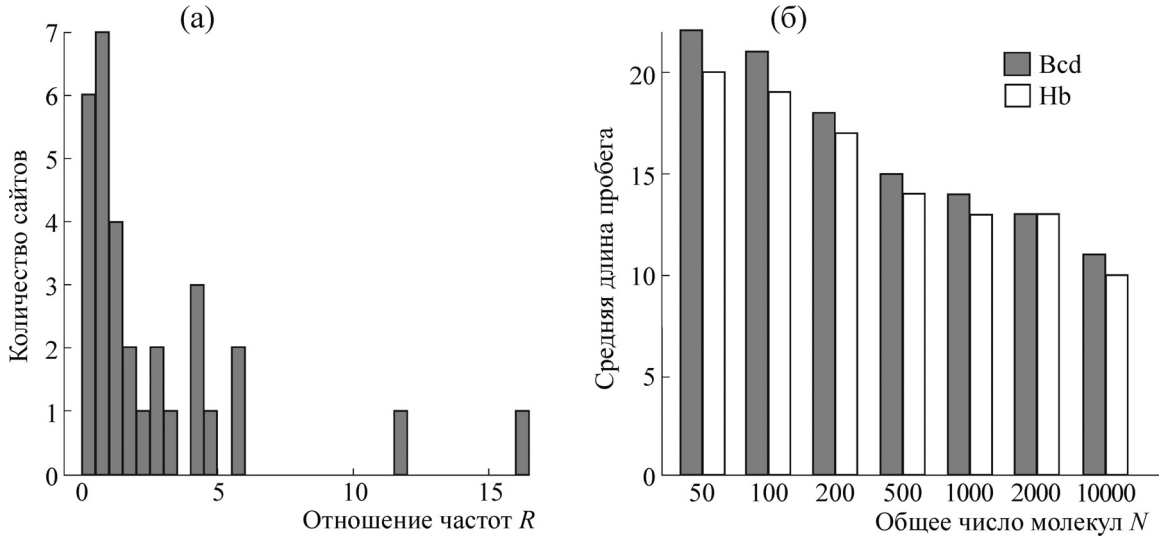

Рис. 3. Биофизические характеристики репрессии и 1D-движения на ДНК, рассчитанные в результате 10000 запусков программы. (а) – Распределение значений отношения R между частотами активного и неактивного состояния для 31 сайта связывания Hb (все сайты Hb на одной из нитей ДНК в регуляторной районе). (б) – Средняя длина пробега белков при различных общих количествах N белков в системе.

ской модели были исследованы биофизические характеристики, связанные с эффективностью репрессии и длиной пробега белков на ДНК. Специфичный сайт, занятый ТФ-репрессором (Hb), может находиться в двух состояниях – активном и неактивном. В активном состоянии связанный с ним белок модифицирует локальную окрестность ДНК и, тем самым, реализует механизм короткодействующей репрессии, тогда как в неактивном состоянии репрессии не происходит (см. Материалы и методы, раздел «Механизм репрессии»). Чтобы оценить, насколько эффект репрессии устойчив для каждого отдельного сайта и для всех сайтов-репрессоров в целом, были проведены 10000 расчетов в рамках модели, в ходе которых для каждого сайта-репрессора оценивались частоты активного и неактивного состояний (f_a и f_n соответственно) и вычислялось их отношение $R = f_a/f_n$. Данная характеристика отражает степень участия данного сайта в репрессии: чем больше значение R, тем больше вклад сайта в общую репрессию; значение R, близкое к нулю, означает, что сайт почти не занимается в активной форме и, как следствие, не участвует в репрессии. Полученное в результате вычислений распределение для R показывает, что вероятность нахождения в активном состоянии для сайтов-репрессоров довольно вариабельна (рис. 3а). Бóльшая часть сайтов Hb либо редко активна, либо вероятность находиться в активном состоянии не сильно отличается от вероятности находиться в неактивном состоянии (медианное значение R для 31 сайта равно 1,35). При этом существуют сайты, которые практически всегда участвуют в репрессии, – это сайты, соответствующие длинному хвосту распределения на рис. 3а.

В термодинамических моделях экспрессии генов, включающих механизм короткодействующей репрессии, сила репрессии учитывается с помощью введения в уравнения модели весов, которые выделяют конфигурации энхансера, содержащие связанные белки-репрессоры в активном состоянии [4,8]. Как правило, для упрощения вычислений в рамках этих моделей для весов используются постоянные значения. Исследования стохастической модели энхансера показывают, что более детализированные характеристики репрессии являются принципиально вариабельными. Параметр R или его аналоги могут служить обобщенной характеристикой силы репрессии, а его распределение – возможным приближением для учета того, как эффект репрессии распределен в регуляторном районе гена-мишени.

Другим важным параметром в процессе формирования конфигураций энхансера является длина свободного пробега белка в ходе скольжения по ДНК. Величины среднего пробега белков Bcd и Hb не сильно различаются, но зависят от общего числа молекул (рис. 3б). Большее число белков в нуклеоплазме, очевидно, увеличивает общую интенсивность их неспецифичного взаимодействия с регуляторным районом, что создает эффект массового скопления белков на ДНК, затрудняя движение по ДНК и уменьшая длину пробега. Этот эффект

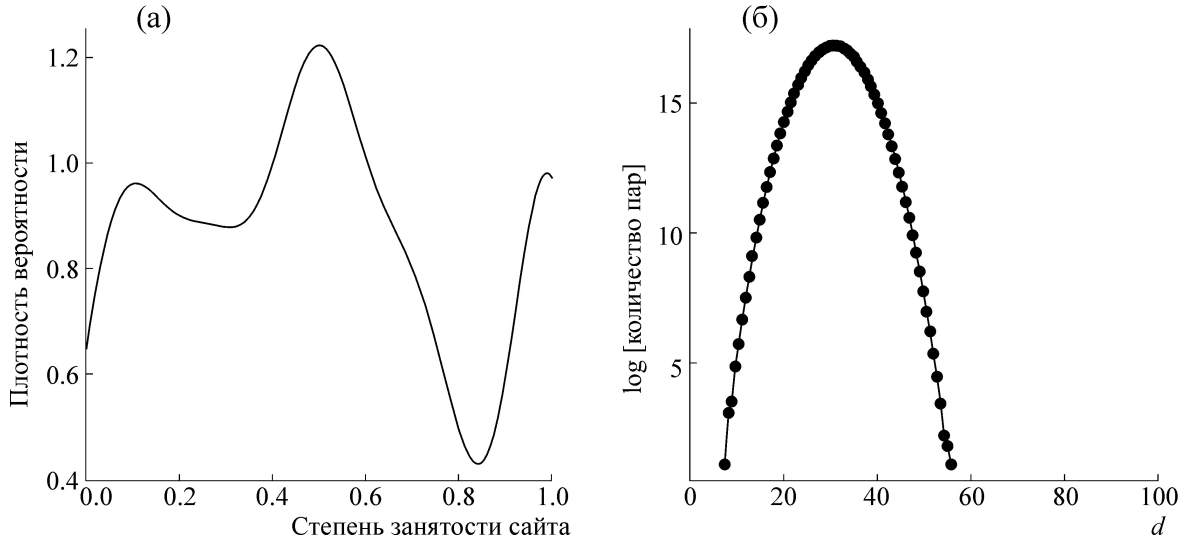

Рис. 4. Вероятность занятости сайтов и вариабельность конфигураций. (а) – Распределение вероятности для степени занятости ССТФ в стационарных конфигурациях регуляторного района, рассчитанное в результате 60000 запусков стохастической модели. Степень занятости сайта определяется как относительная частота тех конфигураций, в которых этот сайт связан с ТФ. (б) – Количество пар конфигураций (в логарифмическом масштабе), отстоящих друг от друга на расстояние Хэмминга *d*.

должен оказывать влияние на уровень экспрессии генов сегментации в ходе раннего развития дрозофилы, поскольку в этот период концентрации ТФ в эмбрионе имеют большие пространственные градиенты, т.е. в разных ядрах эмбриона количество молекул ТФ сильно отличается.

Распределение вероятности связывания для ССТФ. Вероятность появления той или иной конфигурации энхансера зависит от вероятностей связывания для ССТФ, входящих в энхансер. Чтобы оценить распределение вероятности связывания для ССТФ, была проведена серия из 60000 запусков стохастической модели и были вычислены частоты занятости сайтов в квазистационарном состоянии регуляторного района. В рамках модельных допущений полученное распределение имеет три выраженных пика, соответствующих группам сайтов с разной степенью доступности для ТФ (рис. 4а). Наибольшая группа сайтов связывается с ТФ с вероятностью, близкой к 1/2. С точки зрения вычислений это означает, что примерно в половине конфигураций, сформировавшихся к моменту установления стохастического процесса, эти сайты оказывались связанными с ТФ, а в другой половине – свободными. Значительная часть сайтов имеет очень низкую вероятность быть занятыми (крайний левый пик распределения). Такие сайты можно считать практически недоступными для ТФ. Наконец, группа сайтов, соответствующих крайнему правому пику распределения, практически всегда связывается с ТФ. Интересно, что эта группа всегда доступных ССТФ относительно изолирована от остальных двух групп, тогда как отличие между первыми двумя группами сайтов выражено на форме распределения менее характерно.

В результате 60000 запусков, проведенных для вычисления распределения на рис. 4а, получались различающиеся стационарные конфигурации регуляторного района. Чтобы оценить степень вариабельности стационарных конфигураций, были выбраны произвольные 30000 конфигураций и каждая конфигурация из этого множества была параметризована с помощью вектора $\sigma = \{\sigma_i\}$, $i = 1, \ldots, 135$, такого, что $\sigma_i = 1$, если i-й сайт в конфигурации связан с ТФ, и $\sigma_i = 0$, если этот сайт свободен. Получившиеся в расчетах конфигурации были разбиты на всевозможные пары, и были вычислены расстояния Хемминга d между векторами σ конфигураций внутри каждой пары (рис. 4б). Расстояние Хемминга между двумя векторами σ равно количеству различающихся элементов в этих векторах, т.е. числу сайтов, которые имеют разный статус (свободен или занят) в двух конфигурациях. Поэтому d можно выразить в процентах от общего числа сайтов, и чем меньше d, тем более похожи конфигурации. Распределение d в рассмотренном наборе пар конфигураций имеет среднее значение около $d = 35$ (рис. 4б), что соответствует отличию между конфигурациями на 35% (примерно на 47 сайтов из

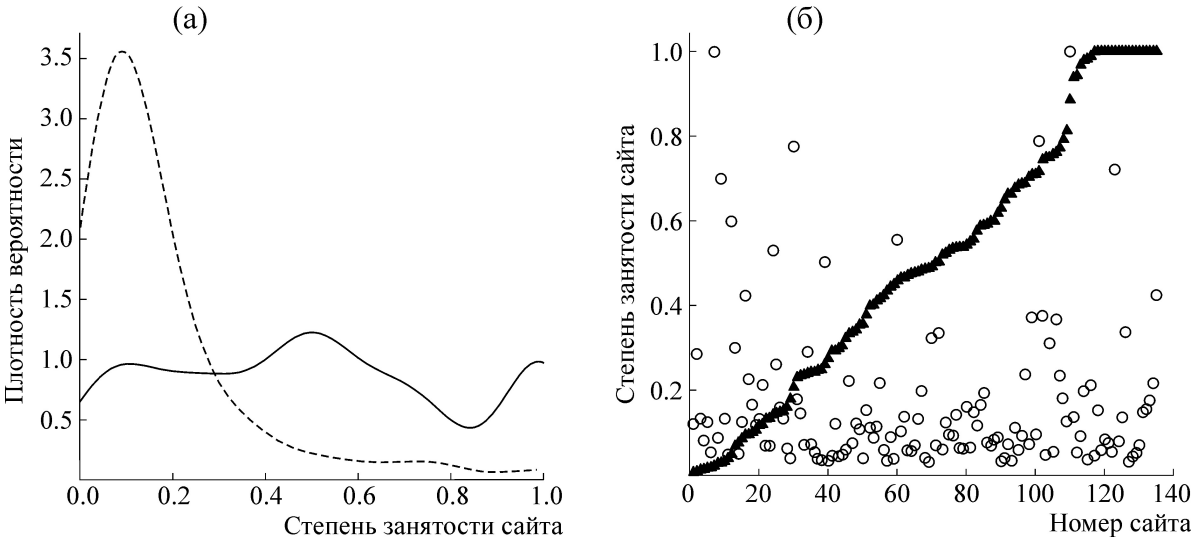

Рис. 5. Сравнение эмпирического распределения вероятности для степени занятости ССТФ с теоретически рассчитанным методами статистической термодинамики. (а) – Плотность вероятности для степени занятости сайтов f_B из формулы (1) (пунктирная кривая) в сравнении с эмпирическим распределением из рис. 4а (сплошная кривая). Для силы репрессии в формуле (1) было принято медианное значение распределения из рис. 3а: $r(j) = 1{,}35$ для всех сайтов-репрессоров. (б) – Значения степени занятости каждого из 135 сайтов в регуляторном районе, рассчитанные из запусков стохастической модели (черные треугольники) и из термодинамической модели (f_B из формулы (1); прозрачные кружки). Сайты упорядочены по возрастанию значений степени занятости из стохастической модели.

135). При этом в целом отличие в конфигурациях варьируется примерно от 5 до 60%.

Для сравнения полученного эмпирического распределения с ожидаемым в рамках допущений термодинамических моделей был произведен расчет степени занятости сайтов с помощью методов статистической термодинамики, описанных в деталях в работах [1,4]. Согласно этим моделям, степень занятости сайта f_{B} вычислялась как отношение суммы статистических весов тех конфигураций, в которых рассматриваемый сайт занят, к общей сумме всех весов (статсумме):

$$f_{\mathrm{B}} = \frac{\sum_b W(b)}{\sum_b W(b) + \sum_u W(u)}, \quad W(\sigma) = \prod_j e^{w(j)} r(j), \tag{1}$$

где $W(\sigma)$ есть статистический вес конфигурации с индексом σ, индексы b в формуле для f_{B} нумеруют конфигурации, в которых рассматриваемый сайт занят, индексы u нумеруют конфигурации, в которых сайт свободен. В формуле для $W(\sigma)$ произведение ведется по всем занятым сайтам j в конфигурации σ, и через $e^{w(j)}$, как и раньше, обозначается энергия присоединения для специфичного сайта с PWM-весом $w(j)$. Параметр $r(j)$ отличен от единицы только для занятых сайтов-репрессоров, находящихся в активном состоянии, т.е. в окрестности которых запрещено связывание других сайтов согласно механизму короткодействующей репрессии. Этот параметр является характеристикой «силы репрессии» в термодинамических моделях.

Применяя формулу (1) ко всем ССТФ в регуляторном районе, было построено распределение степени занятости сайтов (рис. 5а). Сравнение с эмпирическим распределением (полученным в результате стохастических симуляций) указывает на сильное отличие двух распределений. Теоретически рассчитанные значения степени занятости сайтов группируются около значения 0,1. Бóльшая часть сайтов, которые согласно стохастической модели имеют степень занятости, превосходящую 0,5, практически недоступны для связывания ТФ в термодинамической модели (рис. 5а,б).

Результаты сравнения показывают, что учет истории формирования конфигурации энхансера значительно меняет роли сайтов в связывании ТФ по сравнению с предсказаниями из статистической термодинамики. Последние строятся на основе распределения Больцмана для энергий связывания; таким образом, если два сайта находятся примерно в равных условиях (например, оба не пересекаются с другими сайтами и не находятся в окрестности сайтов-репрессоров), то неизбежно сайт с бóльшим сродством к ТФ будет иметь бóльшую степень занятости в термодинамической модели. В стохасти-

ческой модели играют роль дополнительные факторы, которые могут приводить к тому, что сайт с бóльшим весом w будет менее часто заниматься ТФ, чем похожий сайт с меньшим весом. Эти факторы напрямую связаны с важностью процессов неспецифичного связывания ТФ и 1D-движения по ДНК, которые не учитываются в термодинамическом формализме. При учете этих процессов становится важным не только вес самого специфичного сайта, но и вес ДНК-последовательности в его окрестности, поскольку с наибольшей вероятностью белок окажется на специфичном сайте только пройдя эту окрестность. Также важным становится эффект массового скопления белков на ДНК. Если высокоспецифичный сайт находится в центре скопления других сайтов, вероятность достижения этого сайта белком в ходе скольжения по ДНК уменьшается. Наконец, если сайт находится на краю открытой области хроматина, играют роль геометрические ограничения для 1D-движения, поскольку к такому сайту можно пройти по ДНК только с одного конца последовательности. Такие геометрические ограничения, очевидно, будут существенны для регуляторных районов, состоящих из многих модулей, изолированных друг от друга недоступными областями ДНК.

Экспрессию гена-мишени может запустить конфигурация регуляторного района, далекая от стационарного состояния. Чтобы понять, насколько нестационарность процесса формирования конфигурации может влиять на распределение вероятности занятости сайтов, был вычислен аналог распределения из рис. 4а, но для случая, когда в качестве отдельных конфигураций учитывались все промежуточные конфигурации, сформировавшиеся после каждого нового события связывания специфичного ССТФ. Соответствующее распределение качественно совпадает с представленным на рис. 4а, отличаясь лишь высотой отдельных пиков. Этот результат свидетельствует о том, что фактор нестационарности не оказывает большого влияния на различие результатов стохастического моделирования и термодинамической модели.

ЗАКЛЮЧЕНИЕ

В работе представлена стохастическая модель, использованная для моделирования процесса формирования молекулярных конфигураций регуляторного района на примере одного гена-мишени и двух ТФ. Несмотря на упрощающие предположения в рамках модели, она включает в себя все основные механизмы поиска белком своего специфичного сайта связывания и в то же время позволяет за разумное вычислительное время на персональном компьютере выполнить многократные расчеты по «сборке» достаточно больших регуляторных районов. Такой компромисс между степенью детализации модели и ресурсоемкостью вычислений является важным качеством, создающим предпосылки для дальнейшего включения в более сложные модели транскрипционной регуляции в генных сетях.

Расчеты в рамках модели показали значительное расхождение в результатах с приближениями, сделанными в термодинамических моделях экспрессии генов, в которых распределение вероятностей конфигураций основано на распределении Больцмана для энергий связывания ССТФ. В рамках принятых модельных предположений можно сделать вывод, что основным фактором, определяющим это отличие, является не предположение о равновесности системы, сделанное в термодинамических моделях, а отсутствие учета дополнительных факторов, влияющих на вероятность специфичного связывания ТФ с ДНК. Поскольку лимитирующим шагом в общем процессе является неспецифичное связывание, эти факторы связаны с доступностью сайта для белка, совершающего 1D-движение по ДНК.

Отличие эмпирического (рассчитанного в рамках стохастической модели) распределения для вероятности занятости сайта от теоретически рассчитанного в рамках статистической термодинамики априори могло быть двух видов. Поскольку два метода расчета предсказывают разные вероятности для различных конфигураций регуляторного района, можно было ожидать, что либо эмпирическое распределение запрещает много конфигураций, разрешенных в термодинамической теории, либо эмпирическое распределение повышает вероятность многих конфигураций, запрещенных в термодинамической теории. Результаты стохастического моделирования в рамках нашей простой модели свидетельствуют, что скорее верен второй вариант, поскольку два пика в эмпирическом распределении вероятности занятости сайтов соответствуют практически нулевой вероятности в термодинамической модели (см. рис. 5а).

Предложенный метод стохастического моделирования может быть использован как первый шаг для обобщения термодинамических моделей экспрессии генов. Такая обобщенная модель может состоять из двух этапов. На первом этапе с помощью стохастического моделирования формирования конфигураций регуляторных районов всех генов-мишеней в исследуемой системе можно вычислить более реалистичные распределения для вероятностей занятости всех сайтов, а также для основных биофизических параметров (таких как сила короткодействующей репрессии). Полученная информация может быть использо-

вана на втором этапе для вычисления вероятности транскрипционной активации генов-мишеней примерно в том же ключе, как это делается в термодинамических моделях. Такая двухэтапная модель может быть менее требовательной к вычислительным ресурсам в процессе подгонки модели к экспериментальным данным по экспрессии, нежели полная стохастическая модель экспрессии во всей системе.

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (разработка стохастической модели: грант 13-01-00405-а) и Российского научного фонда (расчеты в рамках модели: грант 14-14-00302).

СПИСОК ЛИТЕРАТУРЫ

1. N. E. Buchler, U. Gerland, and T. Hwa, Proc. Natl. Acad. Sci. USA **100** (9), 5136 (2003).
2. E. Segal, T. Raveh-Sadka, M. Schroeder, et al., Nature **451** (7178), 535 (2008).
3. W. D. Fakhouri, A. Ay, R. Sayal, et al., Mol. Syst. Biol. **6** (1), 341 (2010).
4. X. He, C. Blatti, A. H. Samee, and S. Sinha, PLoS Comput. Biol. **6** (9), e1000935 (2010).
5. A.-R. Kim, C. Martinez, J. Ionides, et al., PLoS Genet. **9** (2), e1003243 (2013).
6. A. H. Samee and S. Sinha, Methods **62** (1) 79 (2013).
7. J. M. Dresch, M. Richards, and A. Ay, Biochim. Biophys. Acta **1829** (9), 946 (2013).
8. K. N. Kozlov, V. Gursky, I. Kulakovskiy, and M. Samsonova, BMC Genomics **15** (Suppl. 12), S6 (2014).
9. M. Sheinman, O. Benichou, Y. Kafri, and R. Voituriez, Rep. Prog. Phys. **75** (2), 026601 (2012).
10. I. Izeddin, V. Récamier, L. Bosanac, et al., Elife **3**, e02230 (2014).
11. J. Chen, Z. Zhang, L. Li, et al., Cell **156** (6), 1274 (2014).
12. D. Normanno, L. Boudarène, C. Dugast-Darzacq, et al., Nat. Commun. **6**, 7357 (2015).
13. P. W. Ingham, Nature **335** (6185), 25 (1988).
14. J. Jaeger, Cell. Mol. Life Sci. **68** (2), 243 (2011).
15. X.-Y. Li, S. Thomas, P. J. Sabo, et al., Genome Biol. **12** (4), R34 (2011).
16. J. Jaeger, S. Surkova, M. Blagov, Nature **430** (6997), 368 (2004).
17. O. G. Berg and P. H. von Hippel, J. Mol. Biol. **193** (4), 723 (1987).
18. D. T. Gillespie, J. Phys. Chem. **81** (25), 2340 (1977).
19. D. N. Arnosti, S. Gray, S. Barolo, et al., EMBO J. **15** (14), 3659 (1996).

A Stochastic Model for the Formation of Enhancer Molecular Configurations

G.M. Demidov*, M.G. Samsonova, and V.V. Gursky** *****

**St. Petersburg Academic University, ul. Khlopina 8/3, St. Petersburg, 194021 Russia*

***Peter the Great St. Petersburg Polytechnic University, ul. Polytechnicheskaya 29, St. Petersburg, 195251 Russia*

****Ioffe Institute, ul. Polytechnicheskaya 26, St. Petersburg, 194021 Russia*

Combinations of free and bound sites for transcription factors in the enhancer, which regulates the transcriptional activity of a target gene, form various molecular configurations. The thermodynamic models for gene expression predict the probability of the target gene transcriptional activation based on the probabilities of the enhancer molecular configurations for this gene. In this work, we present a simple stochastic model for the formation of such configurations that takes into account the non-specific DNA binding of transcription factors, the protein sliding along the DNA, and the specific binding to energetically preferable binding sties. We used this model to study the properties of the molecular configurations formed by the regulatory region of gene *knirps*, expressing during the early *Drosophila* development, and transcription factors Hunchback and Bicoid. Under the model assumptions, we show that the process of transcription factor binding to the regulatory region exhibits three consecutive dynamical stages with respect to the binding rate. The transcription factor binding sites group into three distinct sets with different values of fractional occupancies. The quasi-stationary distribution for the binding site fractional occupancies predicted by the stochastic model essentially differs from the one resulted from the thermodynamic equilibrium approximation and used in the thermodynamic models of gene expression. These findings provide a more detailed view of regulatory regions, which can further be used to improve models of gene expression.

Key words: enhancer, stochastic modeling, thermodynamic model, Drosophila, segmentation genes, knirps, Hunchback, Bicoid