БИОФИЗИКА, 2023, том 68, № 5, с. 911–919

МОЛЕКУЛЯРНАЯ БИОФИЗИКА

УДК 577.22

УТОЧНЕНИЕ ПОЗИЦИЙ НУКЛЕОСОМ ВНУТРИ ОТДЕЛЬНЫХ ГЕНОВ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МОЛЕКУЛЯРНОГО МОДЕЛИРОВАНИЯ И ДАННЫХ MNASE-СЕКВЕНИРОВАНИЯ

© 2023 г. В.А. Васильев*, Д.М. Рябов*, А.К. Шайтан*, Г.А. Армеев*, #

Московский государственный университет имени М.В. Ломоносова, Ленинские горы, 1/12, Москва, 119234, Россия

#E-mail: armeev@intbio.org

Поступила в редакцию 06.05.2023 г.
После доработки 06.05.2023 г.
Принята к публикации 17.05.2023 г.

Организация хроматина играет важную роль в регуляции работы генетического аппарата клетки. Основной единицей упаковки хроматина является нуклеосома, хранящая на себе ДНК длиной около 145 пар нуклеотидов. Упаковка генетического материала и его доступность для ферментов транскрипции и других регуляторных хроматиновых белков зависят от позиции нуклеосом. Для исследования позиций нуклеосом в геноме применяют MNase-секвенирование. Данные MNase-секвенирования позволяют детектировать факт наличия нуклеосом на последовательности, однако их точное позиционирование сложно установить по этим данным. Для уточнения положений нуклеосом необходимо дополнительно фильтровать и обрабатывать данные. В данной работе предлагается комбинированный метод отбора возможных позиций нуклеосом по данным MNase-секвенирования, основанный на геометрическом анализе молекулярных моделей нуклеосомных цепочек. Разработанный алгоритм позволяет эффективно отсеивать недоступные комбинации нуклеосомных цепочек и конформационно запрещенные позиции нуклеосом.

Ключевые слова: хроматин, нуклеосома, позиционирование нуклеосом, молекулярное моделирование.

DOI: 10.31857/S0006302923050101, **EDN:** PHCUCE**:**

Геномы большинства эукариот содержат больше ДНК, чем геномы прокариот. Такое различие можно ожидать, так как для работы более сложно устроенных организмов требуется большее число генов. Однако размер генома не связан напрямую со сложностью организма. Так, многие растения обладают значительно большим по длине геномом, чем человек. Причина различий в размерах геномов между эукариотами и прокариотами кроется в ряде причин. Эукариотические геномы содержат некодирующие участки ДНК, расположенные между и внутри генов. В отличие от прокариот, некоторые эукариотические гены многократно повторяются. Наличие множественных копий генов позволяет значительно повысить уровни экспрессии кодируемых белков. Так, например, кодируются белки-гистоны – одни из самых распространенных белков в клеточном ядре.

В геномах бактерий большая часть ДНК кодирует белки. Например, геном *E. coli* состоит примерно из 5 миллионов н.п. и содержит около 4 тысяч генов, при этом почти 90% ДНК кодирует белковые последовательности. Геном пекарских дрожжей, состоящий из 12 миллионов н.п., примерно в 2.5 раза больше генома *E. coli*. В среднем у пекарских дрожжей на один ген приходится около 2000 н.п. и примерно 70% генома дрожжей занято кодирующими последовательностями белков, которых у дрожжей около 6000. Геномы высших животных (например, человека) устроены еще сложнее и содержат продолжительные области некодирующей ДНК. По современным оценкам только 1.5% из примерно 3 миллиардов н.п. генома человека кодирует последовательности белков [1]. Остальные участки заняты некодирующими последовательностями, которые тем не менее выполняют регуляторные функции. Несмотря на линейный размер, эукариотический геном помещается в клеточном ядре, при этом сохраняя свои функции.

Базовым элементом компактизации хроматина является нуклеосома. Нуклеосомы состоят из восьми гистонов (четыре типа гистонов, формирующих гетеродимеры) и ДНК длиной порядка 145 пар нуклеотидов. Нуклеосомы – симметричные структуры (рис. 1а), при этом ось симметрии (диадная ось нуклеосомы) в них проходит вблизи центра одной из нуклеотидных пар (далее по тексту диадная н.п., на рис. 1а показана черной точкой). Упаковка генетического материала и его до-

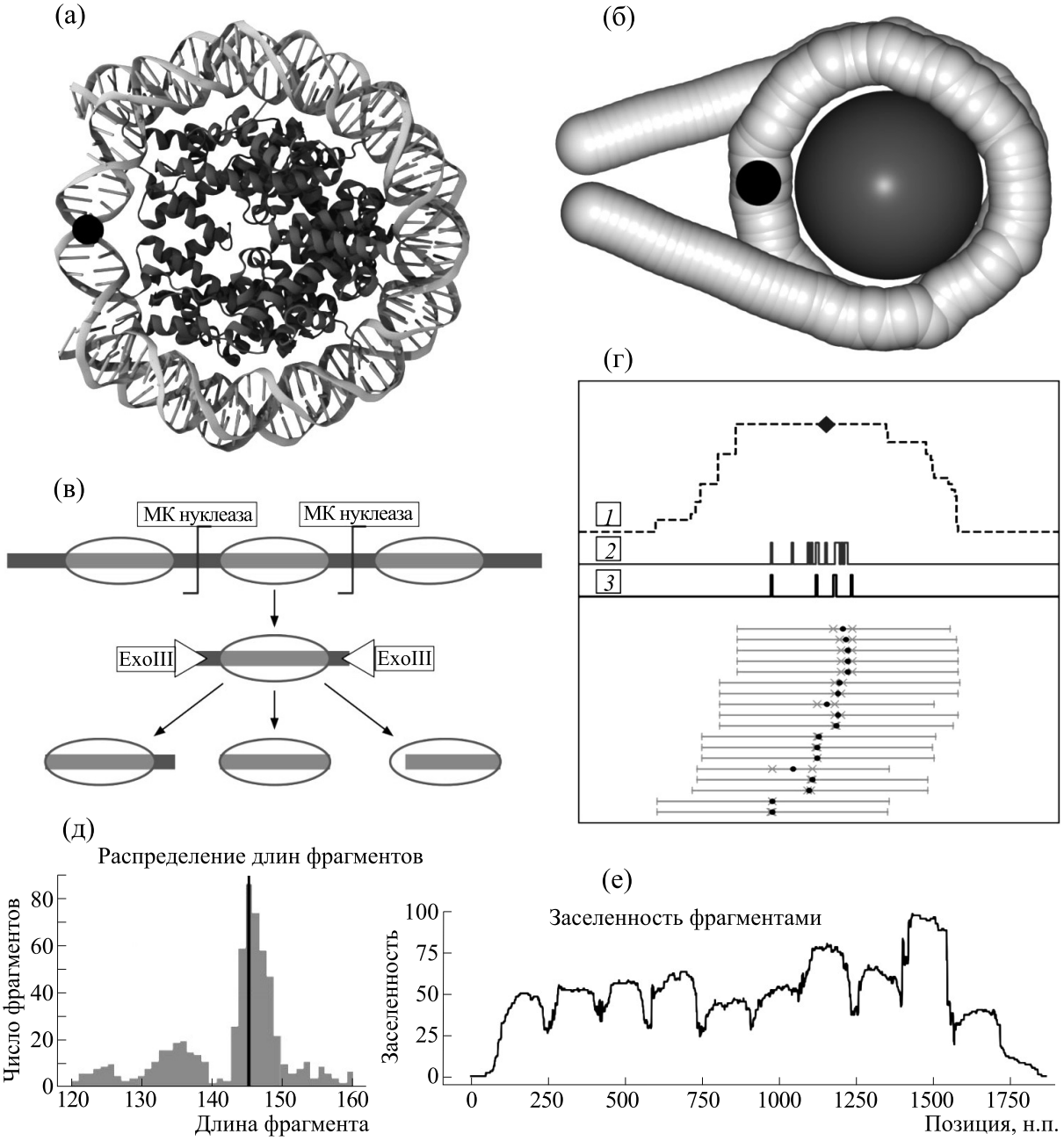

Рис. 1. (а) – Внешний вид структуры нуклеосомы. Визуализация построена на основе модели 3LZ0 из PDB. Белковая часть показана в виде вторичной структуры темно серым цветом, ДНК показана светло-серым. Черной точкой отмечено положение диадной н.п. (б) – Внешний вид огрубленной модели нуклеосомной частицы с линкерными областями. Черной точкой отметено положение диадной н.п. Цветовая схема соответствует рисунку (а). (в) – Схема появления фрагментов разной длины в экспериментах по MNase-секвенированию. (г) – Схема формирования пиков на профилях заселенности нуклеосомами. Отрезки в нижней половине схемы отображают фрагменты нуклеосомной ДНК из MNase-секвенирования, черные точки – положения центров фрагментов, серые крестики – возможные положения диадных н.п. Линия *1* – сумма всех фрагментов, ромбом показано положение центра пика (данное положение часто используют как статистическое среднее расположение нуклеосом); линия *2* – центров сигналов; линия *3* – отфильтрованный профиль положения диадных н.п. (д) – Распределение длин выровненных участков. (е) – Профили заселенности участков генома нуклеосомами.

ступность для ферментов транскрипции и других регуляторных белков хроматина зависят от расположения нуклеосом. Известно, что на стабильность нуклеосом влияет последовательность ДНК. Существуют такие последовательности, на которых нуклеосомы не только стабильны, но и

строго позиционированы [2]. Однако фундаментальные механизмы влияния последовательности ДНК на стабильность и позиционирование нуклеосом неясны. Позиционирование нуклеосом в хроматине обеспечивается не только сродством к последовательности, но и работой транскрипционных факторов, которые активно передвигают нуклеосомы по ДНК (ремоделлеров хроматина). Например, гены, на которых активен ISW1, в среднем содержат нуклеосомы каждые 175 н.п., а гены, регулируемые CHD1, – 160 н.п.[3]

Нуклеосомы в геноме распределены не случайно. Существуют как области с высокой упорядоченностью нуклеосом, так и участки генома, лишенные четко определенной нуклеосомной организации. В работе [4] впервые было показано, что в областях вблизи промоторов *S. cerevisiae* нуклеосомы точно позиционированы. Такое позиционирование связано с высокой транскрипционной активностью и работой транскрипционных факторов [5]. Значимость позиционирования нуклеосом проявляется во многих биологических исследованиях и напрямую влияет на локальную структуру хроматина [6]. Позиционирование нуклеосом может оказывать влияние на процессы регуляции работы ДНК, например, связывание нуклеосом с ДНК может приводить к блокированию сайтов связывания других белков. Интересно, что с течением жизни структура хроматина меняется: снижается уровень экспрессии гистоновых белков, уменьшается заселенность генов нуклеосомами, что в результате приводит к нарушению процессов транскрипции [7].

Как уже было отмечено выше, позиционирование нуклеосом зависит от последовательности ДНК и активности хроматиновых ремоделлеров. Известно, что *in vitro* нуклеосомы способны к спонтанному тепловому передвижению. Однако механизм такого передвижения не до конца ясен, в ряде работ по молекулярному моделированию были показаны начальные этапы перемещения нуклеосом по ДНК [8]. Один из основных методов изучения позиций нуклеосом *in vivo* – MNase seq (от аббревиатуры МНКаза – микрококковая нуклеаза). В этом методе исследуют препарат хроматина, выделенный из клеточной культуры. Такой хроматин далее обрабатывают микрококковой нуклеазой – ферментом, разрезающим фрагменты цепи ДНК, не связанные с белками. Однако данный фермент предпочтительно разрезает АТ-богатые регионы [9]. В результате после обработки получаются фрагменты ДНК, связанные с белками (рис. 1в). В силу селективности эндонуклеазной активности и низкой экзонуклеазной активности, получившиеся фрагменты ДНК могут быть больше по размеру, чем связанные с белками участки. Чтобы уточнить области связывания, образец дополнительно обрабатывают экзонуклеазой III (ExoIII), которая расщепляет не защищенные белками концы ДНК. Полученные фрагменты ДНК очищают от белков и определяют методами высокопроизводительного секвенирования. Получившиеся чтения картируют на геном и фильтруют по длине. Участки длиной порядка 150 н.п. обычно ассоциируют с областями расположения нуклеосом [10]. В ходе дальнейшего анализа рассчитывают распределения длин выравненных участков (рис. 1д) и определяют профили заселенности участков генома нуклеосомами – для каждого нуклеотида определяется число сигналов от участков нуклеосомной ДНК (рис. 1е).

Важно отметить, что полученные сигналы (выравненные на геном участки нуклеосомной ДНК) не позволяют однозначно судить о позиции нуклеосомы. Обработка смесью нуклеаз может приводить к формированию участков ДНК, отличающихся по длине от ожидаемых для нуклеосомы (рис. 1в,д): часть нуклеосом будет обработана недостаточно, а часть – чрезмерно. Таким образом, сигналы в совокупности позволяют уверенно детектировать факт наличия нуклеосом на последовательности, однако их начальное позиционирование неоднозначно. Если считать центры сигналов за диадные н.п. нуклеосом, то для каждого пика заселенности ожидается достаточно большое число возможных положений нуклеосомы (рис. 1г, линия *2*, центры сигналов показаны черными точками). Таким образом, для точного определения положений нуклеосом необходимо дополнительно фильтровать и обрабатывать данные, как например в работах [11, 12].

Эксперименты по определению позиций нуклеосом по большей части получены для культур клеток и содержат сигналы позиционирования для большой совокупности геномов. На итоговых профилях заселенности мы видим результат суперпозиции множества альтернативных вариантов позиционирования, который дополнительно смазан неоднородностью длин сигналов. Определение возможных расположений нуклеосом по таким данным – сложная вычислительная задача. В данной работе мы предлагаем метод отбора возможных позиций нуклеосом по данным MNase-секвенирования. Данный метод основан на отборе наиболее вероятных положений нуклеосом по изначальным сигналам с последующей фильтрацией допустимых комбинаций. Созданный нами фильтр основан на геометрическом анализе доступного конформационного пространства для молекулярных моделей нуклеосомных фибрилл. Мы применяем данный метод для поиска доступных положений нуклеосом на ряде генов *S. cerevisae* по данным эксперимента MNase-секвенирования.

Таблица 1. Характеристики исследуемых участков генома *S. cerevisae*

Открытая рамка считывания (ОРС)	Координата начала ОРС	Координата конца ОРС	Число возможных цепочек нуклеосом	Число возможных позиций нуклеосом	Число стерически возможных позиций нуклеосом	Коэффициент корреляции модельного профиля с экспериментом
YJR046W	522048	523912	154688	101	92	0.76
YOR066W	449436	451375	60843	101	93	0.49
YLR177W	511054	512990	84870	92	65	0.48
YFL041W	49139	51007	51543	89	80	0.51
YHL019C	67731	69548	22664	85	72	0.65
YPR155C	835563	837413	42262	94	82	0.46

МАТЕРИАЛЫ И МЕТОДЫ

Обработка геномных данных и выбор областей для моделирования. Для разработки и апробирования метода использовали набор данных из работы [13]. Из базы архива секвенирований SRA [14] был загружен набор чтений для эксперимента по MNase-секвенированию хроматина *S. cerevisae*, идентификатор эксперимента SRR1802189. Чтения были обработаны и картированы на геном дрожжей (версия сборки генома GCF_000146045.2) с помощью программы Bowtie2 [15]. Полученные сигналы были отфильтрованы по стандартному протоколу программы. Для дальнейшего анализа были выбраны отображения сигналов на (+) цепи ДНК. Для исследования были выбраны шесть случайных генов длиной менее 2000 н.п. При выборе участков для моделирования использовали следующие критерии: сигнал MNase-секвенирования определен от более чем 400 нуклеосом, распределение длин сигналов имеет максимум в 145 н.п., для данных генов имеется аннотация уровнем экспрессии. Аннотация уровнем экспрессии была взята из эксперимента [16]. Список исследуемых генов и их геномные координаты приведены в табл. 1. В результате данной обработки были получены сигналы позиционирования нуклеосом внутри исследуемых генов.

Моделирование цепочек нуклеосом. Для моделирования цепочек нуклеосом использовали огрубленный геометрический подход. В таком подходе каждая пара нуклеотидов и каждое белковое ядро нуклеосомы задается сферическими частицами разных диаметров (рис. 1б). В данном подходе описание геометрии фибриллы проводили во внутренних параметрах ДНК: взаимные ориентации н.п. относительно друга задавались при помощи шести параметров (shift, slide, rise, tilt, roll, twist). Позиции нуклеосом записывали как отступ от диадной н.п. В качестве шаблона для создания нуклеосомных цепочек использовали структуру 3LZ0 [17] (рис. 1а) из банка данных PDB. Для этой структуры были рассчитаны параметры ДНК и относительное положение белкового ядра нуклеосомы при помощи программного пакета PyNAmod. Создание модели цепочки нуклеосом происходило в несколько этапов: создавался набор параметров для ДНК требуемой длины в B-форме; в местах положений нуклеосом параметры ДНК заменялись параметрами для нуклеосомы; проводилась конвертация параметров ДНК в реальное пространство; рассчитывались координаты центров нуклеосом.

Таким образом, данный подход позволяет соединять нуклеосомы прямыми участками двухцепочечной ДНК для дальнейшего расчета их характеристик. В частности, для цепочек нуклеосом рассчитывали количество внутренних стерических перекрываний. Для пар нуклеотидов был взят радиус 3.5 Å, а для белковых ядер нуклеосом – 32 Å. Для каждой пары частиц рассчитывали перекрывание: если расстояние между частицами меньше суммы их заданных радиусов, то принимается, что частицы перекрываются, а соответствующая цепочка нуклеосом отвергается. Для доступных конформаций нуклеосомных цепочек рассчитывали радиус инерции по центрам нуклеосом.

Для проведения всех расчетов применяли программный пакет PyNAMod, для ускорения численной математики использовали библиотеки Numpy и транслятор Numba.

Для поиска доступных цепочек нуклеосом были созданы моделей всех возможных комбинаций цепочек из пяти нуклеосом и длиной соединяющей ДНК от 0 до 100 н.п. Всего было создано 10^8 комбинаций. Расчеты проводили в парал-

лельном режиме с использованием 80 процессорных ядер. Длина свободного участка ДНК на концах цепочек нуклеосом составляла 50 н.п. Цепочки, в которых наблюдались стерические перекрывания, отбрасывались. Затем было рассчитано отношение числа доступных конформаций к общему числу комбинационно возможных цепочек нуклеосом. Для анализа цепочек с шестью и семью нуклеосомами были рассчитаны конфигурации с максимальной длиной линкерных областей 40, так как полный расчет на доступных мощностях занял бы более трех лет.

Моделирование позиций нуклеосом в гене. Для предсказания возможных позиций нуклеосом в гене из всех сигналов MNase секвенирования учитывали длину сигнала. В случае если длина фрагмента соответствовала длине нуклеосомной ДНК (145 пар нуклеотидов), предполагалось однозначное определение диадной н.п. Если фрагмент отличался по длине от нуклеосомной ДНК, предполагалось наличие двух возможных диадных н.п. (рис. 1г). Более короткие сигналы могут получаться по причине того, что фермент расщепил часть нуклеосомной ДНК с одного из концов. Аналогично два возможных положения диадной н.п. возникает в том случае, когда фрагмент длиннее 145 н.п. Из полученных возможных позиций диадных н.п. рассчитывали профиль вероятности обнаружения диадной н.п. по последовательности гена. Согласно этому профилю, для каждого сигнала выбиралась только одна позиция диадной н.п. В случае равной вероятности для центров диад такой сигнал отбрасывался. Для последующего анализа были выбраны позиции диадных н.п., которые были обнаружены в двух и более сигналах.

По отобранным позициям диадных н.п. был построен направленный граф, в вершинах которого находятся кандидаты в позиции нуклеосом, а связи присваиваются для нуклеосом, находящихся не ближе 145 н.п. и не дальше 215 н.п. (среднее расстояние между нуклеосомами для *S. cerevisae* 165 н.п. [18]). Все возможные конформации фибрилл гена можно представить, как путь в этом графе. В качестве начальных вершин путей выбирались вершины, в которые не приходят связи и они находятся на расстоянии не более 250 н.п. от начала гена. Аналогично, конечные вершины – это такие вершины, из которых не начинается связь, а нуклеосомы находятся на расстоянии не более 250 н.п. от конца гена. Из всех начальных вершин находятся все возможные пути до каждой конечной вершины. При построении пути применяли фильтр по допустимым комбинациям длин линкерных областей: при добавлении каждой вершины в путь, если длина пути с новой вершиной больше или равна пяти, проверяется, что комбинация четырех последних длин линкерных областей в пути стерически возможна (см. пункт «Моделирование цепочек нуклеосом»).

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Точное позиционирование важно с геометрической точки зрения, так как сдвиг нуклеосомы на 1 н.п. приводит к небольшому смещению (3.3 Å), но значительному повороту (34.3°). Выровненные участки из секвенирования – сигналы позиционирования нуклеосом – не позволяют однозначно определить позицию диадных н.п. нуклеосом. Обработка смесью нуклеаз приводит к формированию участков ДНК, отличающихся по длине от нуклеосомной (рис. 1в,д): часть нуклеосом будет обработана недостаточно, а часть – чрезмерно. Число коротких и длинных сигналов значительно меньше числа нормальных сигналов (рис. 1д). Для того чтобы участок ДНК был атакован нуклеазой и стал короче, ДНК должна «открутиться» от нуклеосомы. Вероятность одновременного откручивания нуклеосомы с двух концов ниже, чем вероятность одностороннего откручивания, по этой причине для всех коротких сигналов мы предполагали две возможные позиции диадной н.п., соответствующие откручиванию ДНК либо с одной, либо с другой стороны. Аналогично, мы предположили, что более длинные участки ДНК содержат две возможные позиции. Из профилей встречаемости диадных н.п. для каждого сигнала отбирали наиболее распространенную. Из рис. 1г видно, что если считать за позиции диадных н.п. центры сигналов (черные точки), возникает большое количество дополнительных вариантов позиционирования (линия *2*), однако при использовании предположения, описанного выше, число возможных позиций значительно сокращается (линия *3*). Такая процедура отбора позиций нуклеосом позволяет сократить число возможных кандидатов в два-три раза (в зависимости от гена). В других работах для уточнения позиций нуклеосом анализируют не изначальные сигналы, а профили заселенности нуклеосом. Такие профили обычно обрабатывают оконными фильтрами для локальной нормировки и сглаживания. На обработанном профиле заселенности находят все пики, каждый из которых рассматривают как предположительный центр нуклеосомной ДНК. Затем определяют кластеры возможных пиков – группы таких пиков на расстоянии менее 147 н.п. Далее при анализе выбираются возможные комбинации позиций таким образом, чтобы комбинации нуклеосом формировали максимально схожие с изначальными профили заселенности [11]. Однако такой подход основан на усреднении сигналов по множеству клеток и позволяет определять лишь средние позиции нуклеосом в гене.

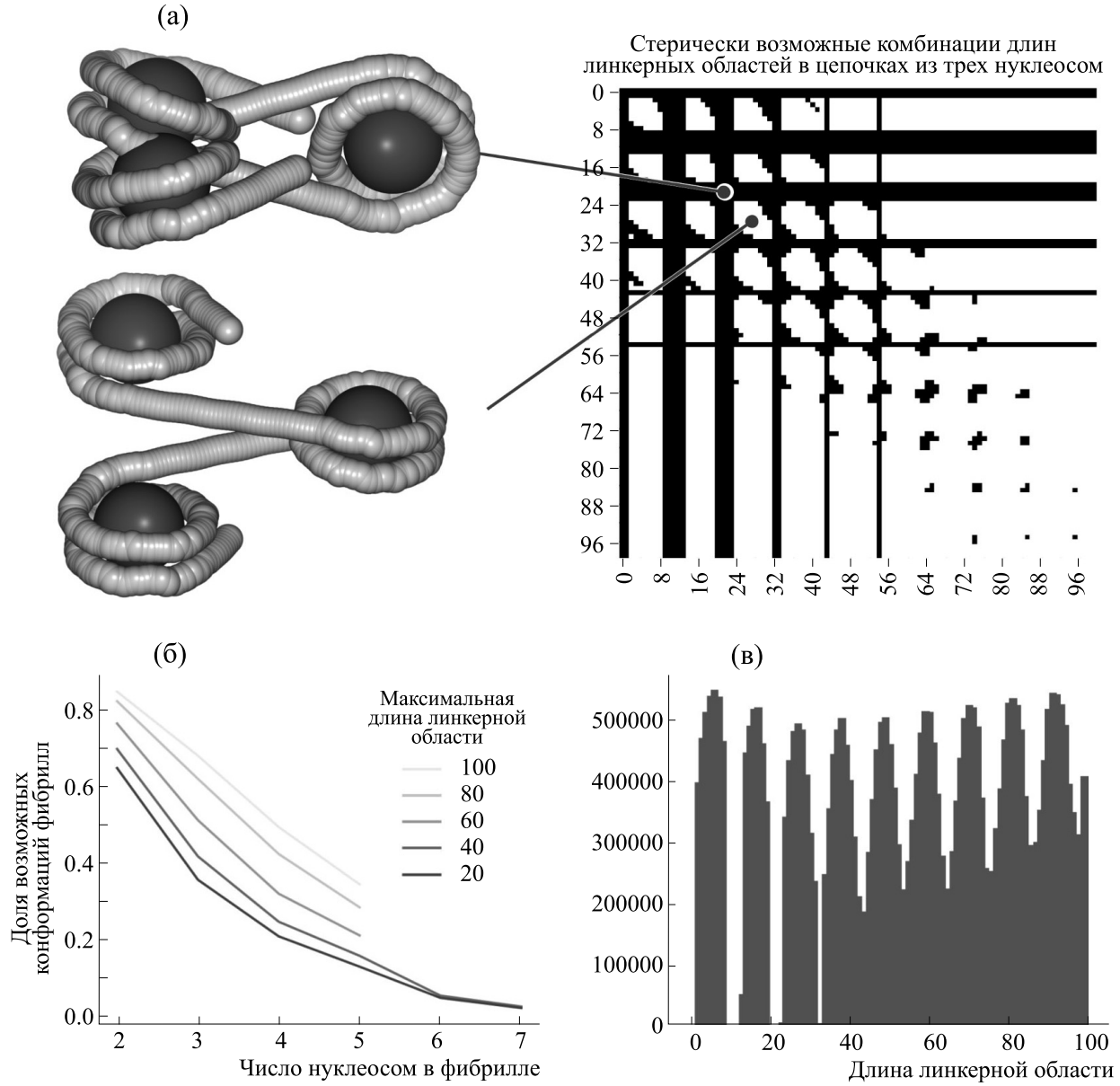

Рис. 2. (а) – Иллюстрация сканирования пространства доступных цепочек из трех нуклеосом. На тепловой карте черным показаны заслоненные конфигурации, белым – разрешенные комбинации линкерных областей ДНК. Слева от карты показаны примеры заслоненных и разрешенных цепочек. (б) – График зависимости доли доступного конфигурационного пространства от числа нуклеосом в цепочке. (в) – Распределение длин линкерных областей ДНК во всех допустимых цепочках из пяти нуклеосом.

Геном *S. cerevisae* отличается сравнительно небольшим размером и высокой долей содержания кодирующих последовательностей. В среднем на один ген у дрожжей приходится порядка 2000 н.п. Для генов со строгим позиционированием нуклеосом средняя длина линкерных областей ДНК составляет порядка 20 н.п. [18] Таким образом, на один ген приходится порядка 10–11 нуклеосом. Если позиция каждой нуклеосомы определена неточно, возможное число комбинаций таких нуклеосом невероятно велико. Однако не любая комбинация позиций нуклеосом возможна геометрически, нуклеосомы могут перекрываться (рис. 2а). Для того чтобы создать фильтр для отбора цепочек нуклеосом, мы рассчитали стерические перекрывания для цепочек длиной от двух до пяти нуклеосом с длиной линкерных областей ДНК от 0 до 100 н.п. и для цепочек длиной 7 и 8 для линкерных областей до 40 н.п. На рис. 2а видно, что ряд комбинаций линкерных областей приводит к появлению заслоненных структур, а на тепловой карте видна периодичность. Такая

периодичность хорошо согласуется с литературными данными. Для регулярных нуклеосомных фибрилл доступен ряд конфигураций, образующий две группы спиральных структур с характерными длинами линкерных областей $10N$ и $10N+5$ [19–21]. Для всех комбинаций фибрилл из 5 нуклеосом чаще всего встречаются ликеры длиной $10N+5$ (рис. 2в). Интересно, что с ростом длины нуклеосомных цепочек доля допустимых комбинаций длин линкерных областей значительно падает, причем для малых длин линкерных областей она приближается к единицам процентов (рис. 2б). Таким образом, для плотных цепочек нуклеосом доступно значительно меньшее конфигурационное пространство чем для цепочек с большим межнуклеосомным расстоянием. Опираясь на данное наблюдение, мы создали алгоритм отбора конфигураций нуклеосомных цепочек произвольной длины по экспериментальным данным.

Мы провели поиск доступных конфигураций цепочек нуклеосом для сигналов позиционирования нуклеосом для шести генов *S. cerevisae*. Разработанный нами подход основан на последовательном переборе возможных комбинаций сигналов нуклеосом на ДНК. Однако доступное пространство комбинаций нуклеосом на гене длиной 2000 н.п. слишком велико для прямого перебора. Так, после обработки сигналов MNase-секвенирования для гена TAH11 (открытая рамка считывания YJR046W) мы отфильтровали 101 возможную позицию нуклеосомных н.п. Исходя из карты заселенности сигналов (рис. 1е), на данном гене находится 10 нуклеосом. Для такого числа сигналов, число сочетаний из 10 нуклеосом составляет порядка 10^{16} комбинаций, однако нуклеосомные сигналы не могут перекрываться. Исходя из числа сигналов, для каждого пика заселенности нуклеосомами ожидается порядка 10 вариантов расположения нуклеосом, что эквивалентно 10^{10} возможных конфигураций цепочек Такое число позиций весьма затруднительно исследовать прямым перебором. Для того чтобы не рассчитывать заведомо недоступные конфигурации нуклеосомных цепочек мы отбрасывали конформацию, как только в ней встречалась запрещенная комбинация позиций для пяти нуклеосом. В итоге для гена TAH11 было обнаружено 154688 возможных комбинаций позиций, что на три порядка меньше изначальной оценки. Интересно, что в результате работы данного алгоритма девять возможных позиций нуклеосом не вошли ни в одну цепочку. Для остальных обработанных генов также была отфильтрована большая часть нуклеосомных цепочек (табл. 1) и отсеяно от 8 до 27 возможных позиций нуклеосом. Все отсеянные позиции встречались не более чем в двух сигналах MNase-секвенирования и, вероятно, относятся к соседним с ними позициям нуклеосом. Для всех доступных моделей был рассчитан радиус инерции от нуклеосомных частиц, который изменялся в пределах от 100 до 250 Å, а полученные модели значительно отличались геометрически (рис. 3б,в). Мы не обнаружили зависимости между определенными параметрами для нуклеосомных цепочек (число доступных конформаций, радиус инерции) и уровнем экспрессии гена. Однако, так как эксперимент по определению уровня экспрессии и эксперимент по картированию нуклеосом были проведены в разных условиях и на разных клеточных культурах, для них могут быть характерны разные состояния хроматина.

Интересно, что полученный модельный профиль заселенности нуклеосом не только качественно совпадает с экспериментальным, но и схож с ним в деталях, о чем свидетельствует высокий коэффициент корреляции Пирсона (рис. 3а, коэффициенты корреляции представлены в табл. 1). Таким образом, тонкая структура профилей в эксперименте является не шумом, а следствием наложения сигналов от нуклеосом из разных клеток. Следовательно, подходы, основанные на сглаживании профилей заселенности нуклеосом с последующим поиском пиков, приводят к потере сигнала точного позиционирования. Для профилей с ярко выраженными десятью пиками заселенности нуклеосом были обнаружены возможные цепочки из 9 и 11 нуклеосом. Доля таких цепочек сравнительно невелика (порядка 5%), однако данные цепочки содержат сигналы с высокой интенсивностью из экспериментальных данных. Интересно, что в модельных профилях заселенности нуклеосомами для всех генов (в отличие от экспериментальных) высоты пиков были равны между собой. Высота пика в эксперименте не связана прямо с представленностью нуклеосомы в геноме, а зависит от времени обработки ферментами и сродства нуклеаз к последовательности линкерных областей ДНК между нуклеосомами [22].

Разработанный нами подход содержит ряд ограничений. В данной модели мы не учитываем гибкость ДНК. Такой подход может приводить к чрезмерной фильтрации конфигураций. Однако, учитывая длину персистентности ДНК порядка 160 н.п. [23] и малую длину линкерных областей ДНК в генах дрожжей, для реализации заслоненных конфигураций потребуется значительно изогнуть ДНК. Также при поиске возможных конфигураций нуклеосомных цепочек, мы ограничивали максимальную длину линкерной области 70 н.п., что позволяет применять подход только для плотно заселенных нуклеосомами участков хроматина.

Разработанный нами алгоритм позволяет эффективно отсеивать недоступные комбинации

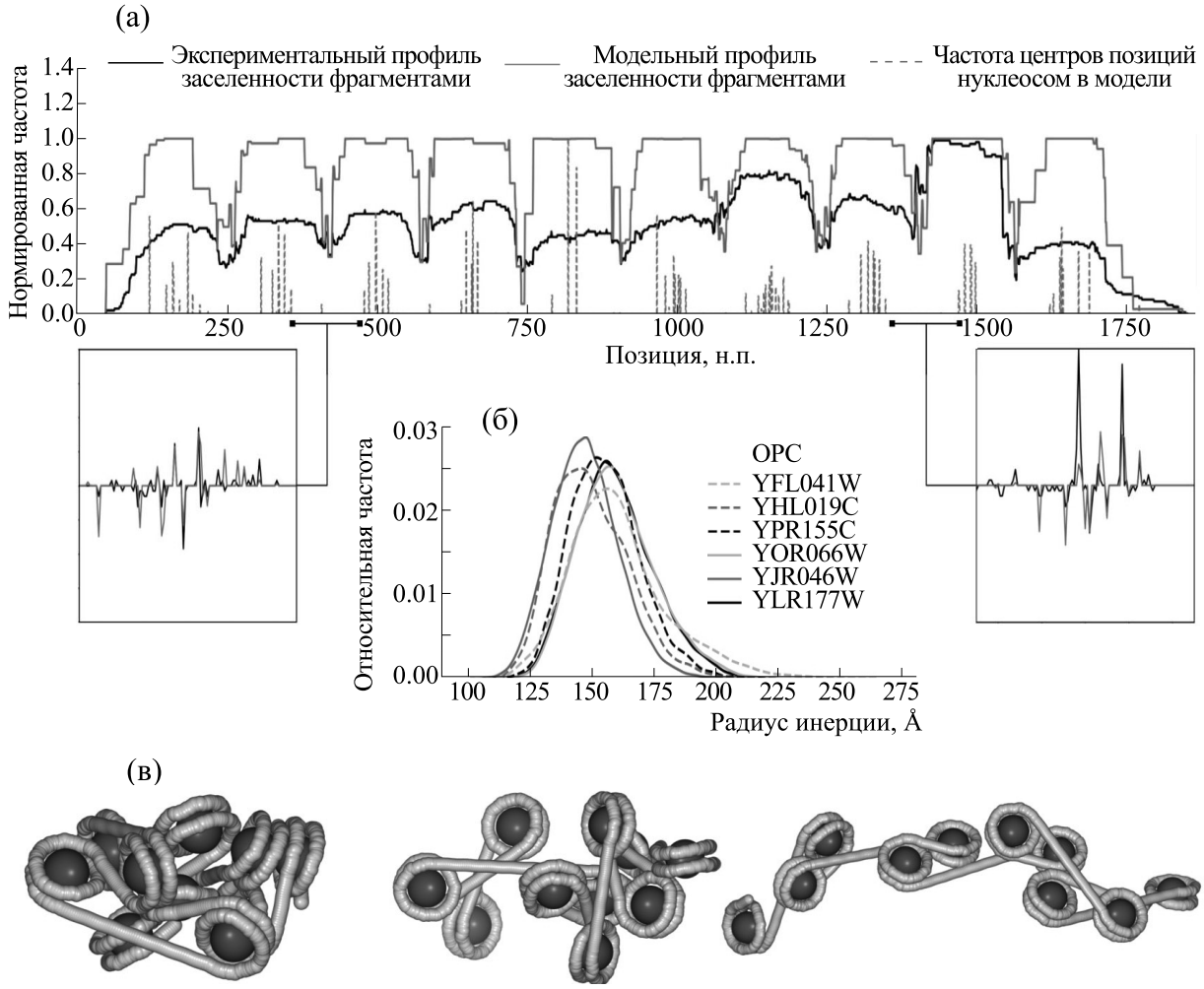

Рис. 3. (а) – Сравнение экспериментального и модельного профилей заселенности фрагментами нуклеосом. Показан ген TAH11 (открытая рамка считывания YJR046W). На врезках показаны дифференциальные профили заселенности. (б) – Распределение радиусов инерции модельных цепочек нуклеосом от разных генов пекарских дрожжей. (в) – Визуализация фибрилл с минимальным, медианным и максимальным радиусами инерции.

нуклеосомных цепочек и конформационно запрещенные экспериментальные сигналы. Однако после такой обработки в результате остаются десятки и сотни тысяч доступных комбинаций нуклеосом. Большое число возможных вариантов может быть объяснено тем, что в экспериментах по MNase-секвенированию наблюдается суперпозиция нуклеосом из клеточной культуры. Конкретное количество геномов, подверженных анализу можно оценить исходя из оптической плотности культуры клеток и объема образца, в эксперименте [13] анализу подвергалось порядка 10^8 клеток. Очевидно, что при таком количестве клеток и случайном расположении нуклеосом в каждой из них профили заселенности были бы равномерными. Учитывая, что на каждый ген приходится порядка 500 сигналов, состояния хроматина, характерные для небольших популяций клеток, будут подавлены доминантными состояниями.

В данной работе предложен и апробирован комбинированный метод отбора возможных нуклеосом по данным MNase-секвенирования и молекулярного моделирования. Разработанный алгоритм позволяет отфильтровывать недоступные комбинации нуклеосом и их позиции. Предложенный алгоритм позволяет уточнять позиции нуклеосом на отдельных генах, однако его можно адаптировать для обработки сигнала от кластеров нуклеосом на полном геноме.

БЛАГОДАРНОСТИ

Работа выполнена с использованием оборудования Центра коллективного пользования сверхвысокопроизводительными вычислительными ресурсами МГУ имени М.В. Ломоносова.

ИСТОЧНИКИ ФИНАНСИРОВАНИЯ

Работа выполнена при финансовой поддержке Российского научного фонда (грант № 21-74-00033).

КОНФЛИКТ ИНТЕРЕСОВ

Авторы заявляют об отсутствии конфликта интересов.

СОБЛЮДЕНИЕ ЭТИЧЕСКИХ СТАНДАРТОВ

Настоящая работа не содержит описания каких-либо исследований с использованием людей и животных в качестве объектов.

СПИСОК ЛИТЕРАТУРЫ

1. G. S. Omenn, Mol. Cell. Proteomics, **20**, 100062 (2021).
2. P. T. Lowary and J. Widom, J. Mol. Biol., **276** (1), 19 (1998).
3. J. Ocampo, et al., Nucl/ Acids Res., **44** (10), 4625 (2016).
4. G.-C. Yuan, et al., Science, **309** (5734), 626 (2005).
5. W. Lee, et al., Nature Genet., **39** (10), 1235 (2007).
6. D. S. Saxton and J. Rine, Proc. Natl. Acad. USA, **117** (44), 27493 (2020).
7. J. Feser and J. Tyler, FEBS Lett., **585** (13), 2041 (2011).
8. G. A. Armeev, et al., Nature Commun., **12** (1), 2387 (2021).
9. C. Dingwall, G. P. Lomonossoff, and R. A. Laskey, Nucl. Acids Res., **9** (12), 2659 (1981).
10. T.-H. S. Hsieh, et al., Cell, **162** (1), 108 (2015).
11. R. Schöpflin, et al., Bioinformatics, **29** (19), 2380 (2013).
12. X. Zhou, et al., eLife, **5**, e16970 (2016).
13. H. A. Cole, et al., Nucl. Acids Res., **44** (2), 573 (2016).
14. R. Leinonen, H. Sugawara, and M. Shumway, Nucl. Acids Res., **39** (Database issue), D19 (2011).
15. B. Langmead and S. L. Salzberg, Nature Methods, **9** (4), 357 (2012).
16. K. Waern and M. Snyder, G3: Genes, Genomes, Genetics, **3** (2), 343 (2013).
17. D. Vasudevan, E. Y. D. Chua, and C. A. Davey, J. Mol. Biol., **403** (1), 1 (2010).
18. T. Tsukiyama, et al., Genes Dev., **13** (6), 686 (1999).
19. N. Kepper, et al., Biophys. J., **95** (8), 3692 (2008).
20. D. Norouzi, et al., AIMS Biophys., **2** (4), 613 (2015).
21. V. B. Zhurkin and D. Norouzi, Biophys. J., **120** (4), 577 (2021).
22. R. V. Chereji, T. D. Bryson, and S. Henikoff, Genome Biol., **20** (1), 198 (2019).
23. J. S. Mitchell, et al., J. Chem. Theory Comput., **13** (4), 1539 (2017).

Updating Nucleosome Positions within Individual Genes Using Molecular Modeling Methods and MNase Sequencing Data

V.A. Vasilev*, D.M. Ryabov*, A.K. Shaytan*, and G.A. Armeev*

Lomonosov Moscow State University, Leninskie Gory 1/12, Moscow, 119234 Russia

Organization of chromatin plays an important role in regulating the genetic machinery of the cell. The basic unit of chromatin packaging is a nucleosome, which harbors DNA of about 145 base pairs in length. The packaging of genetic material and its accessibility to transcription enzymes and other regulatory chromatin proteins depends on the positions of nucleosomes. MNase sequencing is used to examine nucleosome positions in a genome. MNase sequencing data are sufficient for detecting the presence of nucleosomes on the sequence, but a determination of the precise locations of nucleosomes can be problematic. Accurate determination of nucleosome positions requires additional data filtering and processing. In this study, using MNase sequencing data, a combined method based on geometric analysis of nucleosome chain molecular models is proposed for selecting possible nucleosome positions. The developed algorithm efficiently eliminates inaccessible nucleosome chain combinations and conformationally prohibited nucleosome positions.

Keywords: chromatin, nucleosome, nucleosome positioning, molecular modeling