Московский государственный университет имени
М.В.Ломоносова
Биологический факультет
Кафедра биоинженерии

Выпускная квалификационная работа бакалавра
Тема:
Уточнение позиций нуклеосом на ДНК методами молекулярного моделирования

выполнил студент IV курса
Васильев Вениамин Андреевич
Руководитель:
к. ф.-м. н. Армеев Григорий Алексеевич

Москва
2023

Содержание

Введение

ДНК эукариот хранится в виде компактной структуры - хроматина. Основная структурная единица хроматина - нуклеосома. Нуклеосомы представляют комплекс из 8 белков гистонов 4 типов и ДНК. Нуклеосомы важны не только для хранения ДНК, но и для регуляции транскрипции генома. Локальное увеличение числа нуклеосом на гене приводит к снижению активности экспрессии или ее полному прекращению. При потере нуклеосом транскрипция с гена становится более активной. Для регуляции активности генов также важно позиционирование нуклеосом.

Основной метод, применяемый для определения позиций нуклеосом - MNase-seq, который был разработан для изучения позиций нуклеосом в культуре клеток. Также этот метод не позволяет однозначно определить положение нуклеосомы, а лишь указывает на область расположения нуклеосомы с погрешностью в несколько пар нуклеотидов, в связи с чем возникает необходимость в уточнении их положения. В данной работе представлен метод обработки данных, полученных с помощью MNase-seq, позволяющий предположить возможные наборы комбинаций позиций нуклеосом в клетке.

Цели и задачи

Цель данной работы - разработать метод уточнения позиций нуклеосом на ДНК с помощью молекулярного моделирования.

Задачи:

1) Разработать программные модули, рассчитывающие геометрические структуры ДНК и проверяющие наличие геометрических перекрываний в фибриллах ДНК.
2) Проанализировать пространство конформаций возможных коротких нуклеосомных фибрилл.
3) Разработать фильтр возможных позиций нуклеосом, разработать алгоритм выбора стерически возможных комбинаций позиций.

Обзор литературы

1. Методы определения позиций нуклеосом в геноме.

По положению нуклеосом на геноме гистоновые октамеры бывают строго расположенными - такие нуклеосомы последовательно находятся в одном и том же месте в геноме в популяции клеток. Обычно строго расположенными называют нуклеосомы, положения которых варьируются не более чем на 10 нуклеотидов, так как это число сопоставимо с ошибкой определения их положения. Это расположение характерно для геномов дрожжей, реже встречается у высших растений и животных [26, 30]. Нестрого расположенными называют нуклеосомы, положение которых меняется от клетки к клетке. Такое расположение часто встречается у высших растений и животных.

Положение нуклеосом относительно друг друга определяется длинами линкерных областей ДНК. Если длины линкерных областей в последовательности нуклеосом равны, эту последовательность называют регулярной. Если в регулярной последовательности нуклеосомы строго расположены относительно определенной точки, такая последовательность называется фазированной [10].

Также определяют вращательную позицию нуклеосом. Она описывает то, с какой частью спирали ДНК обращена к нуклеосоме. Периодичность спирали B-ДНК примерно 10,5 нуклеотидов, поэтому гистоновые комплексы на расстояниях, кратных этому числу, будут иметь одинаковую вращательную позицию. Обычно в малых бороздках, обращенных к нуклеосоме, большое количество A-T пар, так они способствуют повороту ДНК вокруг нуклеосомы [9, 24].

Для исследования трехмерной структуры хроматина используются методы семейства -C: 3C, 4C, 5C Hi-C. В этих методах осуществляется химическое “сшивание” участков ДНК, контактирующих в хроматине. Затем проводится фрагментация ДНК с образованием “сшитых” пар, лигирование ферментов в этих парах с образованием гибридных молекул ДНК, детекция гибридных молекул [14].

Точность метода Hi-C определяется длиной фрагментов, на которые разрезают ДНК, а также зависит от размера областей, выбираемых на стадии обработки данных (шаг биннировния, bin size), и определяемого размерами хромосом, глубины секвенирования [20].

Широко распространенным способом определения положения нуклеосом на геноме является использование микрококковой нуклеазы (MNase-seq). При таком методе из популяции клеток выделяют ДНК, связанную с нуклеосомами, затем добавляют определенную концентрацию этого фермента, при которой он разрежет только линкерную ДНК, после чего остаются фрагменты, защищенные от расщепления белковыми комплексами. В некоторых модификациях также используется экзонуклеаза III, расщепляющая концевые части полученных фрагментов ДНК, что увеличивает точность метода. В данном методе важно точно подбирать как концентрацию ферментов, так и время обработки хроматина, так как при слишком малом времени будет определена только малая доля нуклеосом с низкой точностью, слишком продолжительное время приведет к излишнему расщеплению ДНК, что также снизит точность определения нуклеосом. Полученные фрагменты секвенируют, соотносят с геномом. По полученным результатам строят плотности вероятности обнаружения нуклеосом возле пары нуклеотидов [7]. Если полученные участки ДНК от разных клеток совпадают, нуклеосома строго расположена (Рисунок 1A). Если участки равноудалены друг от друга, но не совпадают - это

регулярная последовательность с плохо расположенными нуклеосомами (Рисунок 1B). Если положения нуклеосом совпадают и линкерные участки между ними равны - это фазированная последовательность (Рисунок 1C) [10].

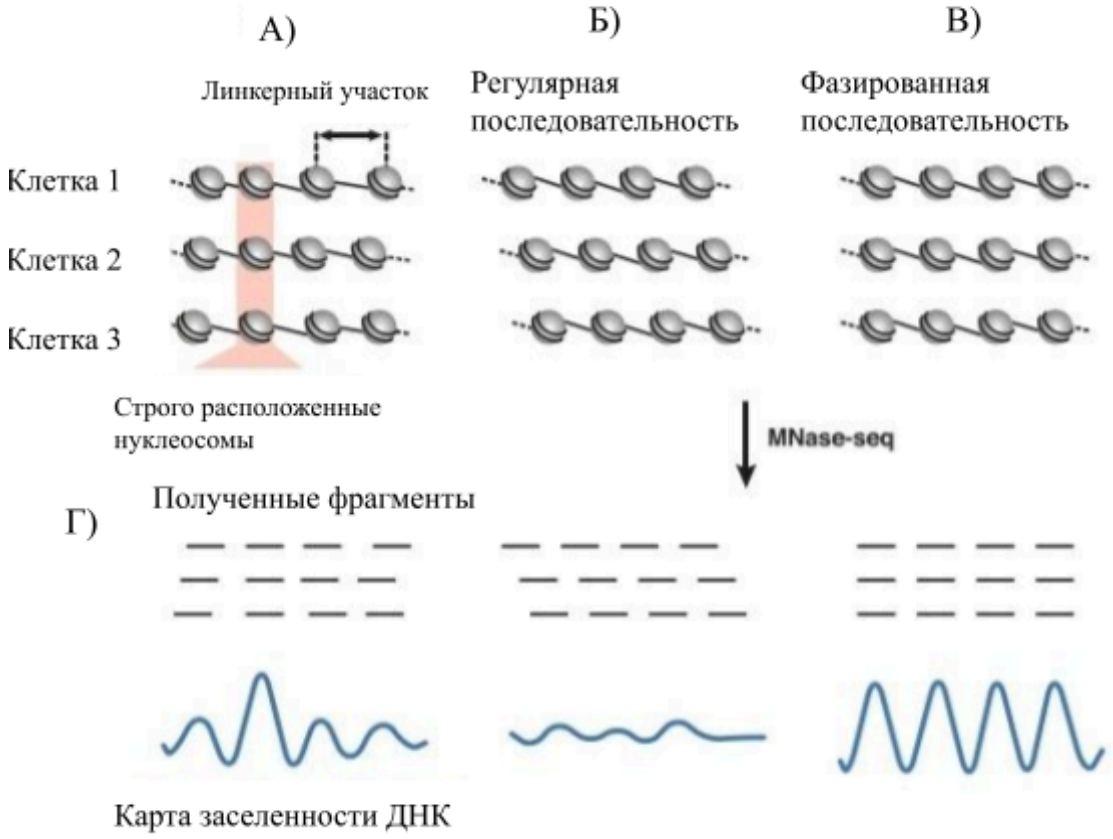

Рисунок 1. Метод MNase-seq. Примеры нестрого расположенных (А), регулярных (Б) и фазированных (В) последовательностей в группах клеток. Г - получаемые из последовательностей фрагменты и карты заселенности ДНК. Адаптировано из [10].

Совмещение методов MNase-seq и Hi-C - Micro-C - позволяет оценивать частоту взаимодействия отдельных нуклеосом друг с другом, позволяя получить разрешение супрануклеосомного уровня [7].

Важно, что методы -С являются статистическими, приблизительно оценивающие частоту взаимодействия фрагментов ДНК, поэтому необходима биоинформатическая обработка полученных данных. [18]Также этот метод плохо определяет положение нуклеосом в плохо расположенных или нерегулярных участках, так как в полученных распределениях нет ярко выраженных пиков [10].

Для определения положения нуклеосом был также разработан ряд независимых от микрококковой нуклеазы методов. К ним относится Hydroxyl-radical-seq, применение которого было показано на дрожжах *in vivo*. В этом методе в гистон H4 вносится мутация, из-за которого в этом белке появляется остаток цистеина, близкий к остову ДНК около нуклеосомной диады. После ковалентного связывания комплекса меди и фенантролина к этому цистеину и добавления дополнительных ионов меди образуются короткоживущие гидроксил радикалы, которые расщепляют ДНК на определенном расстоянии от диады. Это позволяет после секвенирования полученных остатков ДНК с точностью до пары нуклеотидов определить положение диады. Результаты этого метода совпадают с методами, основанными на использовании микрококковой нуклеазы, но дают более высокое разрешение.[5]

2. Факторы, влияющие на позиционирование нуклеосом.

Для предсказания положения нуклеосом используется два основных подхода: позиционирование нуклеосом на основе закономерностей в последовательностях ДНК и на основе различных параметров, полученных при моделировании.

При анализе последовательностей было сформулировано несколько закономерностей в нуклеосомной ДНК. Так, чаще всего RY димеры (R - пуриновые основания: A или G, Y - пиримидиновые основания: T, U или C) наклоняются к малой бороздке, а YR - к большой (Рисунок 2). Кроме того, AT-богатые регионы и CG-богатые регионы также наклоняются к малой и большой бороздке соответственно.

Дальнейшие исследования показали зависимость деформации нуклеотида от контекста: димер CA в YCAR наклоняется в малую бороздку, а в RCAY - в большую; GC наклоняется в малую бороздку в CGCG, а в GGCC - в большую [13].

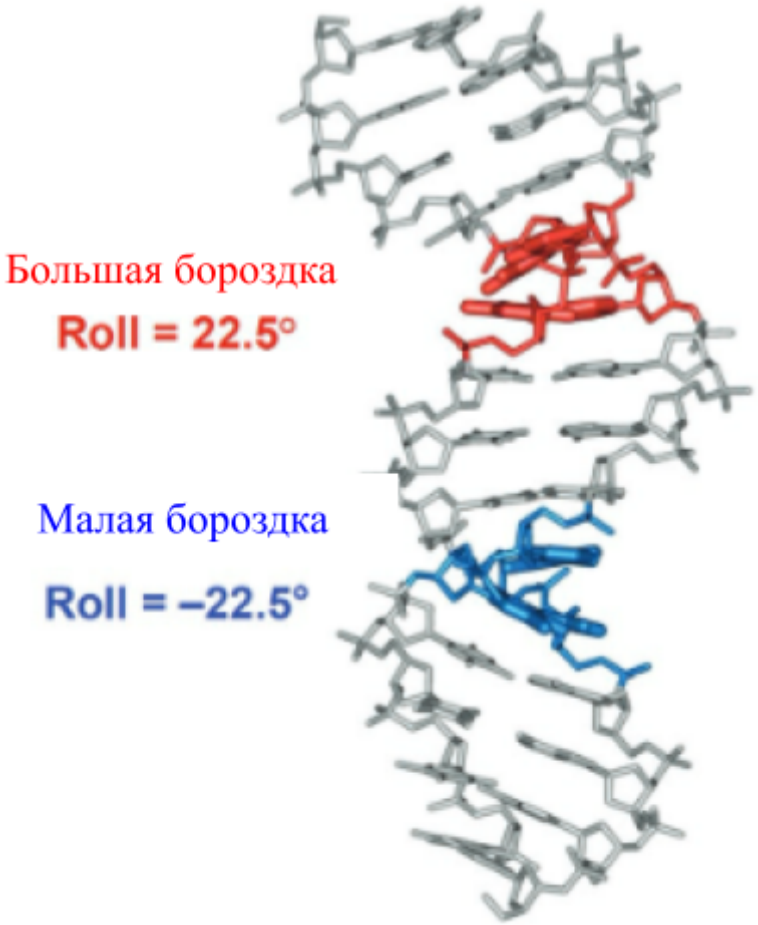

Рисунок 2. Примеры расположения AT (красный) и CG (синий) димеров, повернутых в большую и малую бороздку соответственно, значения углов поворота между ними. Адаптировано из [6].

На основе этих закономерностей строится функция, определяющая вес для каждого шага нуклеотидов в зависимости от наличия определенных мотивов ДНК. Были найдены 3 группы нуклеосом по результатам построения этой функции. У I группы диада расположена в максимуме функции, у II - в локальном максимуме, у III - в минимуме. Разделение объясняется тем, что группа II взаимодействует с гистонами H2A и H2B в сайтах SHL +/-3,5 и +/- 5,5, группа I - в сайтах SHL +/-1,5 (место связывания с тетрамером H3/H4) и +/-4,5 [6].

Второй подход обычно основан на расчетах свободной энергии связывания ДНК с гистонами. Разница в свободной энергии между свободной ДНК и связанной с нуклеосомой ДНК возникает из-за энергии изгиба ДНК вокруг белкового комплекса, энергии взаимодействий ДНК с гистонами, а также энтропийного вклада (например, связанным с высвобождением растворителя из гидратных оболочек вокруг полимеров).

В работе [19] предполагается, что нуклеосомы чаще связываются с ДНК с высоким содержанием CG пар. Это может быть объяснено более высокой энтропией этих пар, так как наблюдаемая более низкая жесткость этих участков не объясняет их энергетическую выгодность, но позволяет им иметь большее количество возможных конформаций. Также было проведено сравнение разных моделей, которое показало, что предсказать более частое связывание нуклеосом с последовательностям с высоким содержанием CG может гибридная модель. В такой модели для расчета энергии нуклеосомной фибриллы константы жестокости для линкерной ДНК определяются по данным из молекулярной динамики, а для нуклеосомной ДНК - по данным из кристаллических структур.

Также существуют алгоритмы для восстановления возможных позиций нуклеосом по данным MNase seq. Для этого по данным прочтений фрагментов ДНК составляют карту плотности заселенности гена, которая показывает, какое число пар чтений картируется на каждую пару нуклеотидов в гене (рисунок 3). Затем карта плотности аппроксимируется с помощью локальных нормальных распределений. Все пики на полученной аппроксимированной карте рассматриваются как возможные положения нуклеосом. Позиции нуклеосом объединяют в кластеры. Кластеры определяются как группа из всех нуклеосомных диад - позиций центров нуклеосомной ДНК, находящихся на расстояние менее 147 пар нуклеотидов от хотя бы одной нуклеосомной диады в кластере. Каждый

кластер исследуется отдельно. В кластерах находят все возможные комбинации позиций нуклеосом, в которых нуклеосомные ДНК разных нуклеосом не перекрываются и добавление новой нуклеосомы приведет к появлению перекрывания. Для полученных комбинаций рассчитываются веса, которые показывают, какой вклад этих позиций нуклеосом в карту плотности заселенности, то есть насколько часто в клетках встречается такая комбинация [23].

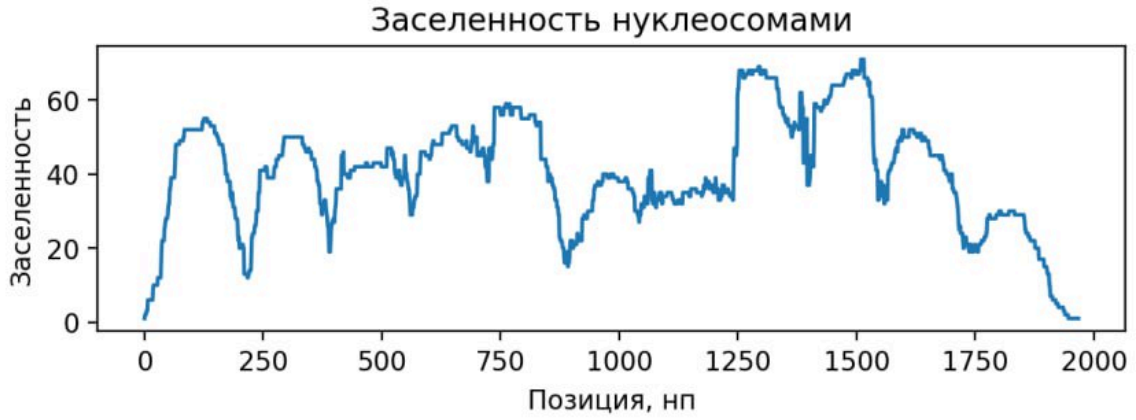

Рисунок 3. Плотность заселенности гена, построенная по данным MNase Seq, взятым из работы [17].

3. Обзор методов моделирования фибрилл нуклеосом.

Для моделирования нуклеосом и их комплексов часто применяют разнообразные методы молекулярного моделирования, в частности методы молекулярной динамики (ссылка на наши работы по МД) и разнообразные методы стохастического моделирования, в том числе для определения трехмерной структуры хроматина (искать по фамилиям разин и чертович)

Для упрощения моделирования фибрилл вместо полноатомной модели используются различные приближения. В одном из наиболее популярных способов описания геометрии ДНК, пара нуклеотидов может описываться 6 параметрами (три угловых - twist, roll, tilt, и три дистанционных - shift, slide, rise), отражающими их смещения относительно друг друга. Для расчета этих параметров для каждого нуклеотида определяется геометрический центр, расстояние от него до начала координат, и ортогональная тройка векторов, определяющая положения нуклеотида. Аналогичные расчеты проводятся для второго нуклеотида в паре, затем рассчитывается среднее между соответствующими векторами, и на основе новых векторов рассчитываются 6 искомых параметров.

Конформационные флуктуации ДНК описываются флуктуациями 6 параметров. Энергия каждого димера аппроксимируется гармонической функцией:

$$E = E_0 + \frac{1}{2}\sum_{i=1}^{6}\sum_{j=1}^{6} f_{ij}\Delta\theta_{\hat{i}}\Delta\theta_{\hat{j}}$$

где $\Delta\theta$ - мгновенные флуктуации параметров, E_0 - минимум энергии, f_{ij} - константы жесткости, препятствующие деформации данного шага. Если объединить константы жесткости в матрицу жесткости димера F, двойная

сумма сводится к $\Theta^T F \Theta$, где элементы Θ - мгновенные флуктуации параметров.

Силовые константы обычно получают из известной (обычно полноатомной) функции энергии, например $f_{ij} = (\partial E/\partial\theta_i\partial\theta_j)$, а попарные ковариации конформационных переменных получают из матрицы, обратной F, например $\langle\Delta\theta_i\Delta\theta_j\rangle k_B T = [F^{-1}]_{ij}$, где k_B - константа Больцмана, T - температура в Кельвинах. В этой модели проводят обратный анализ: ковариации конформационных параметров берут из структур, полученных рентгеноструктурным анализом, вычисляют их ковариацию $\langle\Delta\theta_i\Delta\theta_j\rangle = \langle\theta_i\theta_j\rangle - \langle\theta_i\rangle\langle\theta_j\rangle$ для элементов F^{-1}, затем находят матрицу F инверсией. Однако, Температура для такой матрицы неизвестна, но может быть приблизительно вычислена, например из персистентной длины [4].

В другой модели [3] нуклеосомы, гистоновые хвосты, линкерные ДНК, линкерные гистоны описываются различными приближениями (Рисунок 4):

- Нуклеосома: белковый октамер (за исключением гистоновых хвостов) и ДНК вокруг нее моделируются как одно жесткое тело неправильной формы. Поверхность определяется 300 эффективными зарядами Дебая-Хюккеля, одинаково распределенных по ней. Заряды оптимизируются, чтобы отображать полноатомное электрическое поле на определенном расстоянии от нуклеосомы. Также каждому заряду приписаны исключенные объемы, составляющие за исключенный объем нуклеосомы. 10 гистоновых концов моделируются как выходящие из нуклеосомы последовательности бусинок. Первая такая бусина закреплена к нуклеосоме, их центры соответствуют бета атомам углерода аминокислот, а заряды - сумма зарядов, умноженная на коэффициент, учитывающий ион-зависимое экранирование.

- Линкерная ДНК: аппроксимируется цепью сфер с ион-зависимым зарядом и исключенным объемом для предотвращения перекрывания с другими элементами.
- Линкерные гистоны: моделируется на основе предсказанной структуры H1d крысы. Две сферы, закрепленные относительно друг друга, моделируют С-концевой домен, еще одна сфера представляет N-концевой домен. Каждому линкерному гистону присвоен заряд Дебая-Хюккеля (с учетом ион-зависящего экранирования) и объем. Цепь сфер линкерного гистона расположена на оси диады каждой нуклеосомы.
- Раствор и ионы: вода рассматривается как континуум вокруг нуклеосомы. Экранирование электростатических взаимодействий, связанное с присутствием однозарядных ионов, описано потенциалом Дебая-Хюккеля.

Поворот, перемещение и вращение меняются в соответствии с алгоритмом Монте-Карло. Энергия олигонуклеосомы оценивается как сумма энергий скручивания, изгиба, растяжения линкерной ДНК, растяжения и межмолекулярных изгибов гистоновых хвостов, полной электростатической энергии и вклада исключенных объемов [3].

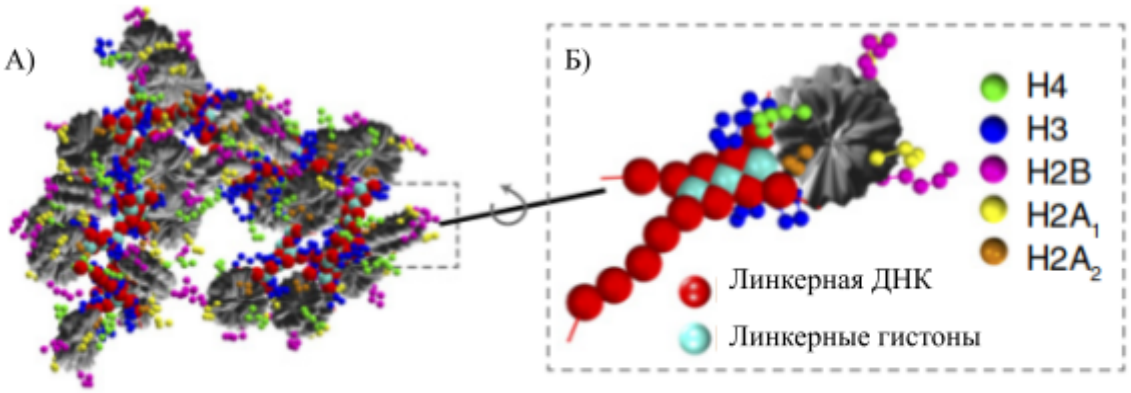

Рисунок 4. Представление модели огрубленной нуклеосомной фибриллы. А - общий вид фибриллы, Б - отдельная нуклеосома (серым цветом) с концевыми доменами гистонов, линкерной областью ДНК и линкерными гистоном. Адаптировано из [3].

4. Распределение нуклеосом по геному, белки, изменяющие положение нуклеосом на геноме.

Выделяют несколько факторов, определяющих положение нуклеосом в геноме.

Последовательности нуклеотидов в ДНК.

Было показано, что разные цепи ДНК имеют разное сродство к нуклеосоме. Это объясняется тем, что свободная энергия, необходимая для закручивания ДНК вокруг гистонового октамера, зависит от последовательности нуклеотидов. Обычно к нуклеосоме обращена малая бороздка, содержат большое количество A/T пар, а малые бороздки, обращенные от нуклеосомы, богаты G/C парами [12].

В короткой цепи ДНК будет несколько пересекающихся последовательностей, к которым будет преимущественно присоединяться нуклеосома. Обычно такие положения смещены на ~10 пар - период повторения спирали (то есть у нуклеосомы будет одно и то же вращательное положение).

Сравнение нуклеосомных карт высокого разрешения, полученных *in vivo* и *in vitro*, показывает что в геноме есть определенные последовательности, с которыми могут связаться нуклеосомы, но распределения гистоновых комплексов по этим позициям отличается *in vivo* и *in vitro*. Это позволяет предположить, что в геноме есть большое количество позиций нуклеосом, из которых клетка может выбирать.

Негистоновые барьеры.

Участки с низким содержанием нуклеосом (nucleosome depleted region, NDR) играют ключевую роль в определении организации хроматина. Предполагается, что негистоновые белки в NDR выступают в

качестве барьера от формирования нуклеосом, вытесняя их к краям региона. Однако на данный момент природа этих белков не выяснена. Вероятно, это должны быть белки, специфично связывающиеся с ДНК, так как NDR формируются в определенных регионах, таких как промоторы, энхансеры, тРНК-кодирующие гены и точки начала репликаций [11].

Транскрипция.

Полимераза II способна проводить транскрипцию на участке, связанном с нуклеосомой, с удалением димера H2A-H2B или, если с ДНК связано большое количество полимераз, полным удалением гистонов с ДНК. Чаще всего, нарушение нуклеосомных фибрилл происходит в наиболее часто транскрибируемых генах [16]. Также в таких генах в среднем короче линкерные участки (например, ~159 пар для 200 самых активных генов дрожжей и ~165 пар для всего генома) [31]. Таким образом, частота транскрипции важный фактор для организации хроматина.

АТФ-зависимые ремоделеры хроматина.

Ремоделеры играют ключевую роль в определении организации хроматина. Существует большое количество таких ферментов, в их функции входит перемещение нуклеосом по ДНК, регуляцию расстояния между нуклеосомами, удаление или перенос нуклеосом, замена гистоновых вариантов.

SWI/SNF и RSC классы ремоделеров могут перемещать и удалять нуклеосомы, определяют длину NDR участков. RSC ферменты в особенности важны для определения позиций нуклеосом возле NDR участка [2]. INO80 ремоделер также влияет на образование NDR участков и позиционировании нуклеосом возле них [29]. ISW1 ферменты меняют длину линкерных участков в генах [27]. У дрожжей взаимодействие этих классов ремоделеров приводит к нерегулярности нуклеосом, так как

положение близких к NDR участку нуклеосом будет зависеть от того, какой из ферментов определяет расстояние между нуклеосомами. Эта нерегулярность влияет на трехмерную структуру хроматина [32].

Таким образом, предполагается, что взаимодействие RSC, INO80 и барьерных комплексов приводит к появлению NDR участков и определяет положение нуклеосом возле таких участков. Эти нуклеосомы становятся референсными при работе ISW1 и CHD1. Предполагается, что ремоделеры переносят нуклеосомы между позициями, определенными последовательностью ДНК.

Репликация ДНК.

Репликация ДНК приводит к нарушению структуры хроматина из-за разрыва водородных связей между двумя цепочками ДНК, прохождения хеликазы и репликационного комплекса. Факторы сборки восстанавливают старые нуклеосомы и собирают новые за репликационной вилкой. У дрожжей ориджин репликации связан с ORC, который влияет на образование NDR участков и позиционированию соседних нуклеосом [22].

Модификации ДНК и гистонов.

Метилирование ДНК влияет на положение нуклеосом у многих эукариот. Периодичные шаблоны метилирования линкерных участков в генах, независимо от уровня транскрипции, вносят вклад в позиционирование нуклеосом. Ковалентные модификации ДНК, такие, как 5-формилцитозин, влияют на подвижность ДНК, стабильность нуклеосом и их позиционирование [21].

Аминокислоты гистоновых хвостов подвергаются посттрансляционным модификациям, таким как ацетилирование, метилирование и фосфорилирование. Эти модификации влияют на

физические свойства нуклеосом, могут отвечать за взаимодействие с ремоделерами. Например, при повреждении ДНК гистоновые модификации становятся сигналом для связывания ремоделерами, которые открывают доступ репарационным факторам к поврежденному участку [28].

Материалы и методы

Программный код был написан на языке Python с использованием библиотек Pandas[33], Numpy[34], Numba[35], Networkx[36]. Разработка программных модулей проводилась как часть разработки программного пакета Pynamod. Для проведения расчетов использовался вычислительный кластер Newton.

1. Анализ геометрических параметров ДНК.

Для анализа фибрилл ДНК использовалось геометрическое описание: положение каждой нуклеотидной пары задается с помощью 3 линейных и 3 геометрических параметров, определяющих расстояние и угол поворота нуклеотидной пары относительно предыдущей пары. Такое описание позволяет независимо менять параметры каждого шага пар нуклеотидов, что упрощает генерацию фибрилл или их моделирование методом Монте Карло. Для описания фибриллы в реальном пространстве каждый последовательный шаг пар нуклеотидов представляется координатной рамкой: набором из трех ортогональных единичных векторов, описывающих поворот шага и вектора, описывающего положения шага в пространстве.

Получение геометрических параметров реализовано в нескольких программах, например, 3DNA [1]. Однако такие программы имеют ряд недостатков: алгоритм их работы не описан полностью в открытом доступе, они не адаптированы для интеграции в другие программные библиотеки. Поэтому в ходе выполнения данной работы был написан модуль, повторяющий алгоритм работы 3DNA. Модуль анализирует полноатомные структуры PDB формата, определяет нуклеотиды и пары между ними, рассчитывает геометрические параметры пар.

2. Анализ геометрического перекрывания фибрилл ДНК.

Для анализа конформаций фибрилл был написан модуль, проверяющий геометрическое перекрывание частиц. В качестве входных данных модуль использует позиции диад нуклеосом и длину фибриллу. Далее генерируется линейная ДНК, которая представляется набором 6 усредненных геометрических параметров для каждого шага в ней. Усредненные значения были взяты из работы [4]. Затем определяются нуклеосомные пары ДНК, для которых значения геометрических параметров заменяются на соответствующие значения, полученные при анализе структуры нуклеосомы 3LZ0. По полученным геометрическим параметрам восстанавливается структура фибриллы в реальном пространстве. Пары нуклеотидов и нуклеосомы представляются сферами с заданными радиусами (3.4 Å для пар нуклеотидов и 32 Å для нуклеосом), описанные с помощью положения центра (Рисунок 5). Затем для всех частиц попарно проверяется наличие геометрического перекрывания.

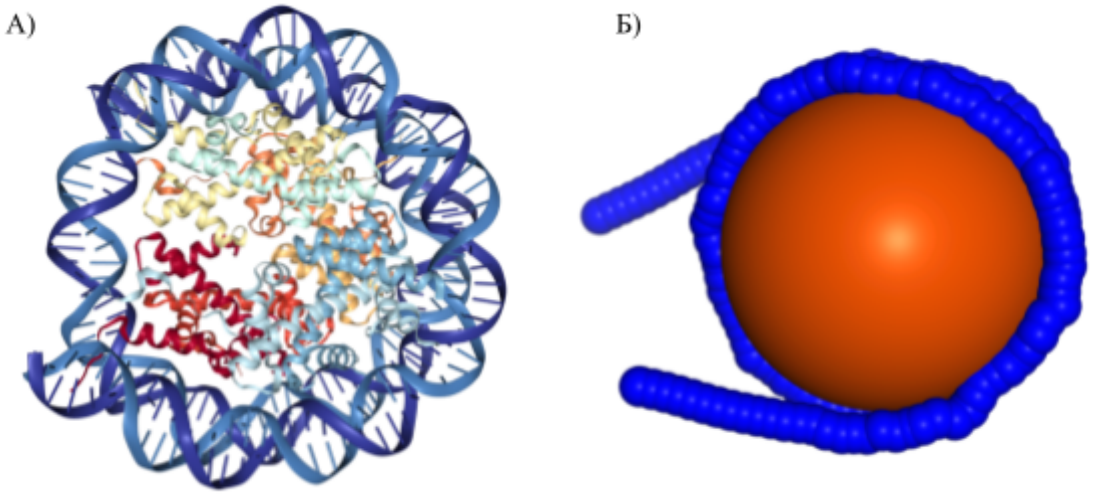

Рисунок 5. Представление структуры нуклеосомы на основение модели 3LZ0 из PDB (А) и в виде огрубленной модели (Б).

3. Изучение пространства доступных конформаций фибрилл.

Для изучения конформационного пространства фибрилл с различным количеством нуклеосом генерировались все возможные фибриллы с длинами линкерных областей до 100 н.п. На концах фибрилл добавлялся свободный участок длиной 50 н.п. Затем в них проверялось наличие геометрических перекрываний. Были проанализированы фибриллы с 3, 4 и 5 нуклеосомами. Всего было проанализировано 10^4, 10^6, 10^8 возможных фибрилл соответственно для разного числа нуклеосом. Изучение фибрилл с большим количеством нуклеосом проводилось для линкерных областей длиной до 40 н.п. Для первичного отбора возможных фибрилл с 6 нуклеосомами ко всем фибриллам с 5 нуклеосомами поочередно добавлялась линкерные области длиной до 40 н.п и еще одна нуклеосома. Из полученных фибрилл с 6 нуклеосомами отбирались те, в которых комбинация последних 5 нуклеосом не имеет перекрывания по данным из моделирования фибрилл с 5 нуклеосомами. Затем для отобранных фибрилл проводился анализ аналогично анализу фибрилл с меньшим числом нуклеосом. Таким же способом с помощью данных по разрешенным конформациям фибрилл с 6 нуклеосомами был проведен анализ фибрилл с 7 нуклеосомами.

4. Анализ возможных конформаций фибриллы ДНК в генах *S.Cerevisae*.

На первом шаге анализа положений нуклеосом в гене определялись возможные позиции диад. В качестве исходных данных использовались прочтения из MNase секвенирования. В случае, если длина прочтения равна длине нуклеосомной ДНК - 145 н.п. - предполагаемое положение нуклеосомной диады совпадает с центром прочтения. В остальных случаях в качестве возможных центров предполагались 2 возможные позиции нуклеосомной диады - 73 пары нуклеотидов, отсчитывая от начала и конца прочтения. После того как из всех прочтений получены центры возможных нуклеосомных диад, для каждого прочтения выбирается та позиция нуклеосомной диады, которая встречается чаще среди всех прочтений. Вторая возможная позиция прочтения отбрасывается. Если обе позиции прочтения встречаются одинаково часто, прочтение отбрасывается. Таким образом, из каждого неотброшенного прочтения получается одна позиция нуклеосомной диад.

Затем, для полученных позиций нуклеосомных диад строится направленный граф, в вершинах которого все позиции нуклеосомных диад, а ребра показывают следующую возможную позицию нуклеосомной диады в гене. Возможные следующие позиции выбираются так, чтобы они были на расстоянии от 145 н.п. до 215 н.п. Путь в таком графе представляет возможный набор позиций нуклеосомных диад в гене (Рисунок 6). В качестве начальных и конечных позиций выбираются такие позиции нуклеосомных диад, которые находятся на расстоянии 250 н.п. от начала или конца гена и не имеют предыдущих (следующих) возможных позиций нуклеосомных диад. В графе находятся все возможные пути от каждой начальной позиции нуклеосомной диады до всех конечных позиций. При поиске применяется фильтр по доступным комбинациям 4 линкеров между

5 нуклеосомами: как только при построении пути встречается запрещенная комбинация, такой путь отбрасывается. Фильтр был взят из предварительного анализа геометрического перекрывания фибрилл с 5 нуклеосомами.

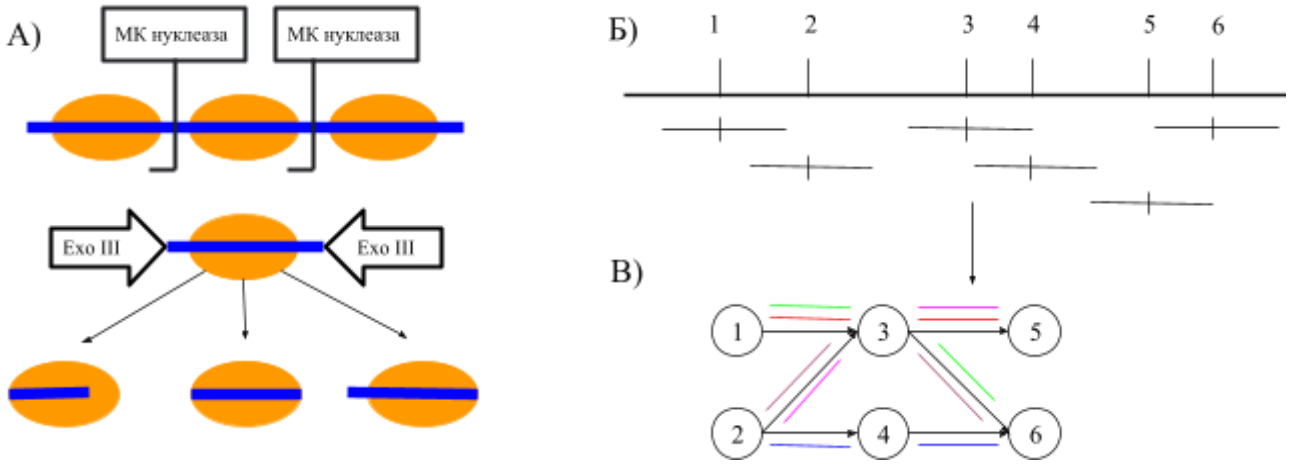

Рисунок 6. А - схема образования прочтений различных длин. Б - Пример отображения возможных позиций нуклеосомных диад из прочтений на участок генома. В - граф, получающийся из примера позиций диад. Все возможные пути, отмеченные разными цветами, показывают возможные комбинации позиций нуклеосом на участке гена.

Результаты работы и обсуждение

Уточнение позиций нуклеосом важно для понимания механизмов регуляция и укладки хроматина. Небольшие изменения позиции могут значительно изменить конформацию хроматина из-за того, что ДНК представляет двойную спираль. Так, смещение нуклеосомы на 1 н.п. приводит к повороту нуклеосомы относительно ДНК на 34.3°. Поэтому важно точное позиционирование нуклеосом, но существующие методы определения позиций *in vivo* имеют погрешности в несколько нуклеотидов. Также в методе MNase Seq большинство прочтений не совпадают по длине с нуклеосомной ДНК (Рисунок 7). Это может объясняться недостаточной обработкой нуклеазами в одних случаях. В других случаях нуклеазы расщепляют нуклеосомную ДНК, отвернутую с одной из сторон от нуклеосомы. В данной работе не рассматривается вариант отворота с двух, так как он менее вероятен, чем отворот с одной стороны. В связи с указанными неточностями методов возникает необходимость в дополнительной обработке данных.

Рисунок 7. Распределение длин прочтений в ген TAH11 (открытая рамка считывания YJR046W).

При создании фильтра по стерически доступным конформациям были проанализированы все возможные комбинации длин линкерных областей до 100 н.п. для фибрилл с 2-5 нуклеосомами и до 40 н.п. для фибрилл с 6-7 нуклеосомами. Полученные данные для фибриллы с 3 нуклеосомами можно представить с помощью тепловой карты. На тепловой карте можно видеть периодичность запрещенных и разрешенных конформаций (Рисунок 8). Это согласуется с литературными данными [8, 15]. Также для фибрилл с 5 нуклеосомами наиболее частые длины линкерных областей - 10N+5 (Рисунок 9 Б). Число доступных конформаций падает при увеличении числа нуклеосом, и для фибрилл с 7 нуклеосомами достигает долей процента. Также видно, что для плотных нуклеосомных фибрилл число доступных конформаций меньше, чем для менее плотно заселенных (Рисунок 9 А).

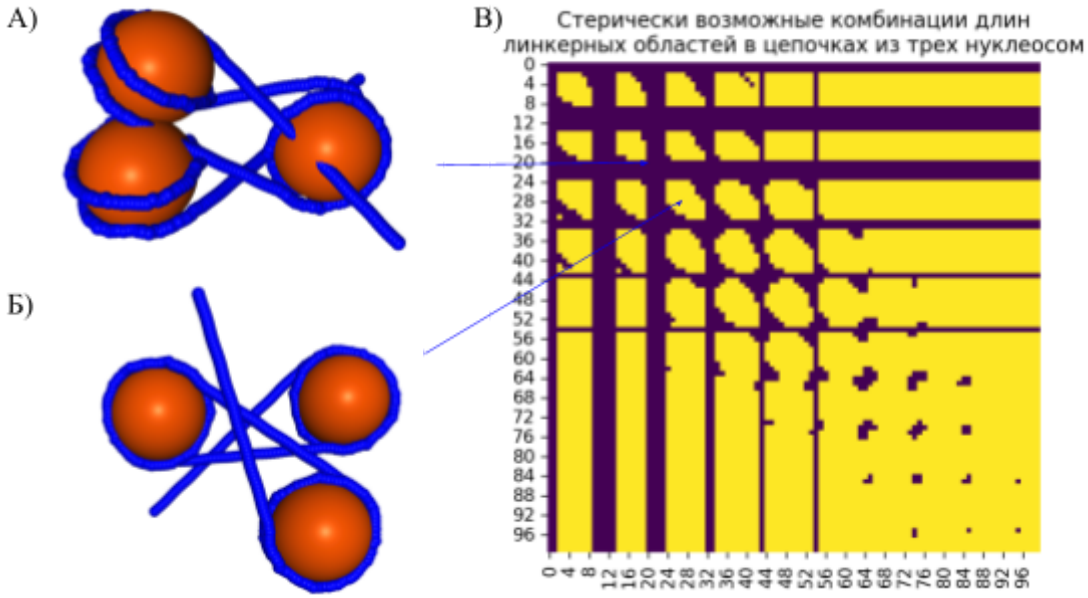

Рисунок 8. Доступные конформации фибриллы с 3 нуклеосомами (В). По осям отложены длины линкерных областей (н.п.). Черным отмечены запрещенные комбинации длин, желтым - доступные. А - пример запрещенной конформации, Б - разрешенной.

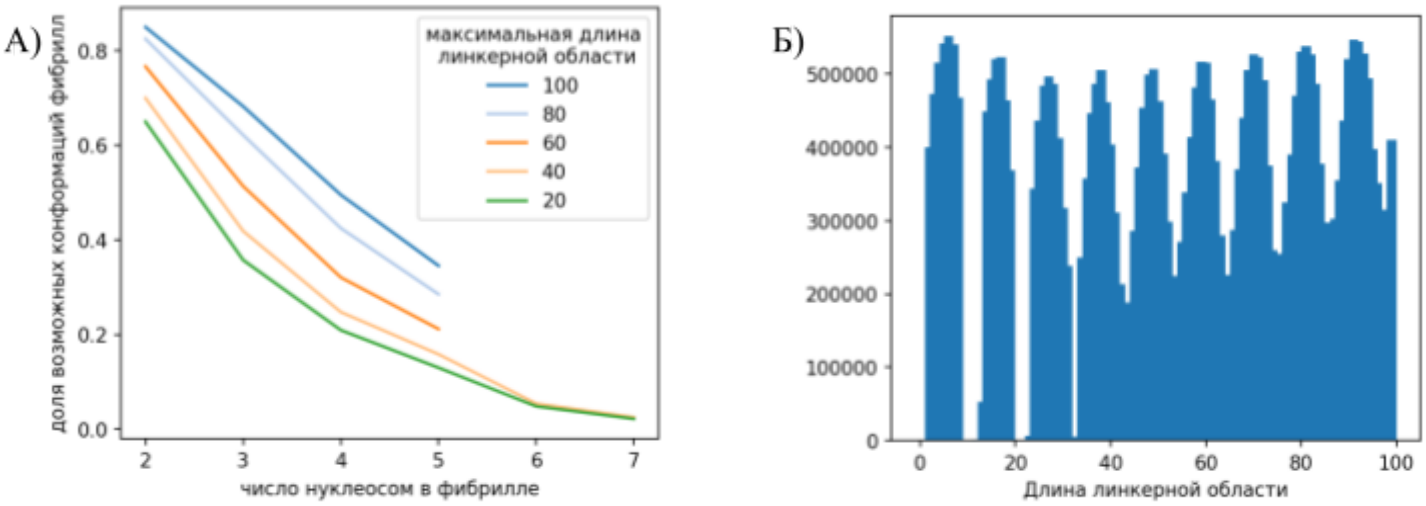

Рисунок 9. А - Зависимость числа доступных конформаций от числа нуклеосом в фибрилле. Б - Распределение длин линкерных областей в возможных фибриллах с 5 нуклеосом.

Мы предложили оригинальную методику выбора возможных позиций нуклеосомных диад по индивидуальным прочтениям, полученным с помощью Mnase Seq(см. методы). Отличие предлагаемого метода в том, что в других алгоритмов обработки позиции нуклеосомных диад выбираются из пиков карты заселенности чтениями. Такой подход позволяет предположить изначально больше возможных позиций нуклеосом, часть из которых теряется при выборе пиков на профиле заселенности. Однако, изначально возникает слишком большое количество возможных позиций нуклеосом. Для отбора меньшего количества позиций для каждого прочтения выбирается только одно возможное положение нуклеосомной диады. Предполагаемая позиция соответствует случаю избыточной или недостаточной обработки нуклеазой фрагмента, и встечается во всех прочтениях чаще, чем другие предполагаемые позиции анализируемого фрагмента. Такой фильтр позволяет сократить число предполагаемых позиций в зависимости от гена в 2-3 раза.

Разработанный алгоритм был применен к 6 генам, прочтения которых взяты из работы [17]. В гене TAH11 (открытая рамка считывания YJR046W) после применения фильтра были получены 101 возможная позиция нуклеосом. Длина данного гена составляет порядка 2000 н.п. нуклеотидов, а среднее расстояние между нуклеосомами - 20 н.п. [25]. Таким образом, предположительное число нуклеосом в гене - 10, а общее число возможных комбинаций позиций составляет порядка 10^{10}, что слишком много для простого перебора. Поэтому нами был использован фильтр по доступным конформациям фибрилл с 5 нуклеосомами. Его применение позволило отобрать 154688 возможных конформаций. При этом 9 позиций нуклеосом не попали ни в одну конформацию. Среди возможных конформаций около 5% содержат 9 или 11 позиций нуклеосом, что подтверждает возможность потери информации при выборе позиций

нуклеосомных диад по пикам на карте заселенности, так как на карте 10 пиков. Для найденных конформаций был рассчитан радиус инерции, его значения меняются от 100 до 250 Å, также конформации сильно отличаются геометрически (Рисунок 10).

По полученным конформациям был построен модельный профиль заселенности. Модельный профиль заселенности совпадает с экспериментальным, что подтверждается высоким коэффициентом корреляции Пирсона (0.76). Это подтверждает то, что вид профиля нуклеосом связан с положением нуклеосом в клетках и возникает в результате их наложения (Рисунок 10 А).

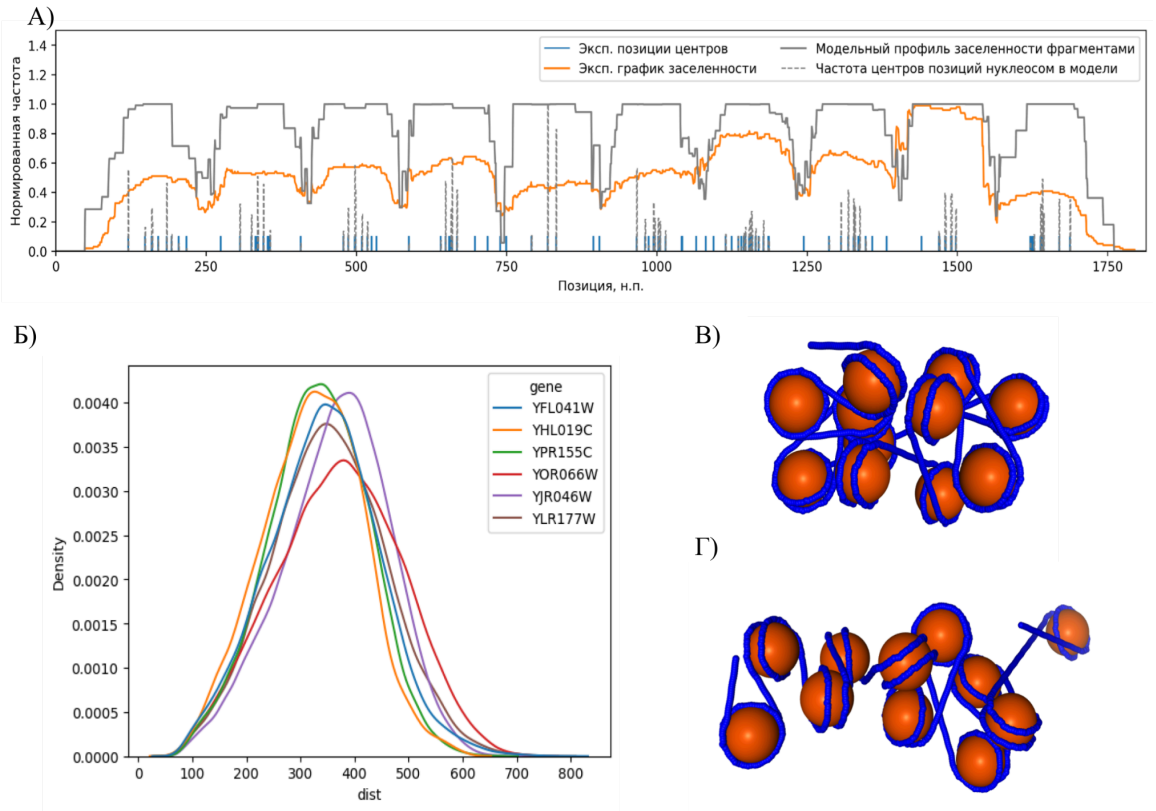

Рисунок 10. А - Сравнение экспериментального и модельного профиля заселенности нуклеосом гена TAH11 (открытая рамка считывания YJR046W). Б - Распределение радиусов инерции выявленных возможных

цепочек нуклеосом. В, Г - примеры возможных цепочек нуклеосом с низким и высоким радиусом гирации соотвественно.

Данные анализа 6 генов представлены в таблице 1. Нами не обнаружено зависимости между радиусами инерции возможных цепочек нуклеосом, их числом и уровнем экспрессии гена. Однако это может быть связано с тем, что эксперименты по определению позиций нуклеосом и измерению уровня экспрессии проводились в разных условиях, при которых позиции нуклеосом в клетках не совпадают.

Таблица 1. Результаты анализа данных по положению нуклеосом участков генома *S.cerevisae.*

Открытая рамка считывания	Коорд. начала ОРС	Коорд. конца ОРС	Число возможных цепочек нуклеосом	Число возможных позиций нуклеосом	Число стерически возможных позиций нуклеосом	Коэфф. 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

Заключение

В данной работе предложен метод получения возможных комбинаций позиций нуклеосом на основе анализа индивидуальных прочтений, полученных с помощью MNase Seq. Метод позволяет эффективно отбирать возможные конформации участков генома. Возникающее большое количество стерически доступных цепочек может быть объяснено тем, что изначальные данные получаются из культуры клеток, в которых могут реализованы разные конформации нуклеосомных фибрилл. Предложенный нами метод был использован на отдельных генах, однако может быть расширен для анализа всего генома.

В работе используется допущение о линейности линкерных областей. При возможности их изгиба с учетом их средней длины порядка 20 нуклеотидов фибрилл будут высокую энергию изгиба, что делает такие фибриллы энергетически менее выгодными. Также разработанный метод подходит только для анализа плотнозаселенных нуклеосомами участков генома, так как было выбрано ограничение 70 н.п. на длину линкерной области.

Также было изучены возможные конформации фибрилл с небольшим числом нуклеосом. Показано, что доля доступных конформаций уменьшается с увеличением числа нуклеосом.

Результаты и выводы

1) Созданы программные модули для геометрического анализа цепочек ДНК по данным структур из pdb, а также для проверки геометрических перекрываний в нуклеосомных фибриллах.
2) Изучены возможные конформации коротких нуклеосомных фибрилл, показано, что:

- Доступное пространство конформаций уменьшается с ростом числа нуклеосом в фибрилле и достигает долей процента;
- В фибриллах можно выявить регулярность: Наиболее часто встречаемые линкеры имеют длину 10N+5.

3) Разработан подход для фильтрации позиций нуклеосом по данным MNase seq. Разработанный метод позволяет предложить возможные комбинации позиций нуклеосом в отдельных генах. Он позволяет значительно снизить пространство возможных взаимных расположений нуклеосом.

Список литературы

1. Lu X.-J., Olson W. K. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures // Nucleic Acids Research. 2003. № 17 (31). С. 5108–5121.
2. Ganguli D. [и др.]. RSC-dependent constructive and destructive interference between opposing arrays of phased nucleosomes in yeast // Genome Research. 2014. № 10 (24). С. 1637–1649.
3. Collepardo-Guevara R., Schlick T. Chromatin fiber polymorphism triggered by variations of DNA linker lengths // Proc Natl Acad Sci U S A. 2014. № 22 (111). С. 8061–6.
4. Olson W. K. [и др.]. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes // Proceedings of the National Academy of Sciences of the United States of America. 1998. № 19 (95). С. 11163–11168.
5. Brogaard K. [и др.]. A map of nucleosome positions in yeast at base-pair resolution // Nature. 2012. № 7404 (486). С. 496–501.
6. Cui F., Zhurkin V. B. Structure-based analysis of DNA sequence patterns guiding nucleosome positioning in vitro // Journal of Biomolecular Structure & Dynamics. 2010. № 6 (27). С. 821–841.
7. Hsieh T.-H. S. [и др.]. Micro-C XL: assaying chromosome conformation from the nucleosome to the entire genome // Nature Methods. 2016. № 12 (13). С. 1009–1011.
8. Norouzi D. [и др.]. Topological diversity of chromatin fibers: Interplay between nucleosome repeat length, DNA linking number and the level of transcription // AIMS biophysics. 2015. № 4 (2). С. 613–629.
9. Albert I. [и др.]. Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome // Nature. 2007. № 7135 (446). С. 572–576.

10. Baldi S. Nucleosome positioning and spacing: from genome-wide maps to single arrays // Essays in Biochemistry. 2019. № 1 (63). C. 5–14.
11. Chereji R. V., Clark D. J. Major Determinants of Nucleosome Positioning // Biophysical Journal. 2018. № 10 (114). C. 2279–2289.
12. Drew H. R., Travers A. A. DNA bending and its relation to nucleosome positioning // Journal of Molecular Biology. 1985. № 4 (186). C. 773–790.
13. Gorin A. A., Zhurkin V. B., Olson W. K. B-DNA twisting correlates with base-pair morphology // Journal of Molecular Biology. 1995. № 1 (247). C. 34–48.
14. Hakim O., Misteli T. SnapShot: Chromosome Conformation Capture // Cell. 2012. № 5 (148). C. 1068.e1-1068.e2.
15. Kepper N. [и др.]. Nucleosome Geometry and Internucleosomal Interactions Control the Chromatin Fiber Conformation // Biophysical Journal. 2008. № 8 (95). C. 3692–3705.
16. Kulaeva O. I., Hsieh F.-K., Studitsky V. M. RNA polymerase complexes cooperate to relieve the nucleosomal barrier and evict histones // Proceedings of the National Academy of Sciences. 2010. № 25 (107). C. 11325–11330.
17. Leinonen R. [и др.]. The sequence read archive // Nucleic Acids Research. 2011. № Database issue (39). C. D19-21.
18. Nagano T. [и др.]. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure // Nature. 2013. № 7469 (502). C. 59–64.
19. Neipel J., Brandani G., Schiessel H. Translational nucleosome positioning: A computational study // Physical Review E. 2020. № 2 (101). C. 022405.
20. Pal K., Forcato M., Ferrari F. Hi-C analysis: from data generation to integration // Biophysical Reviews. 2019. № 1 (11). C. 67–78.
21. Raiber E.-A. [и др.]. 5-Formylcytosine controls nucleosome positioning through covalent histone-DNA interaction // 2017. C. 224444.
22. Ramachandran S., Henikoff S. Replicating Nucleosomes // Science

Advances. 2015. № 7 (1). C. e1500587.
23. Schöpflin R. [и др.]. Modeling nucleosome position distributions from experimental nucleosome positioning maps // Bioinformatics (Oxford, England). 2013. № 19 (29). C. 2380–2386.
24. Trifonov E. N., Sussman J. L. The pitch of chromatin DNA is reflected in its nucleotide sequence. // Proceedings of the National Academy of Sciences. 1980. № 7 (77). C. 3816–3820.
25. Tsukiyama T. [и др.]. Characterization of the Imitation Switch subfamily of ATP-dependent chromatin-remodeling factors in Saccharomyces cerevisiae // Genes & Development. 1999. № 6 (13). C. 686–697.
26. Valouev A. [и др.]. Determinants of nucleosome organization in primary human cells // Nature. 2011. № 7352 (474). C. 516–520.
27. Vary J. C., Fazzio T. G., Tsukiyama T. Assembly of Yeast Chromatin Using ISWI Complexes Chromatin and Chromatin Remodeling Enzymes, Part A / Academic Press, 2003.C. 88–102.
28. Williamson E. A. [и др.]. Chapter 8 - Overview for the Histone Codes for DNA Repair Mechanisms of DNA Repair / под ред. P. W. Doetsch, Academic Press, 2012.C. 207–227.
29. Yen K., Vinayachandran V., Pugh B. F. SWR-C and INO80 Chromatin Remodelers Recognize Nucleosome-free Regions Near +1 Nucleosomes // Cell. 2013. № 6 (154). C. 1246–1256.
30. Zhang T., Zhang W., Jiang J. Genome-Wide Nucleosome Occupancy and Positioning and Their Impact on Gene Expression and Evolution in Plants // Plant Physiology. 2015. № 4 (168). C. 1406–1416.
31. Functional roles of nucleosome stability and dynamics | Briefings in Functional Genomics | Oxford Academic [Электронный ресурс]. URL: https://academic.oup.com/bfg/article/14/1/50/2731572 (дата обращения: 23.05.2023).
32. Mesoscale Modeling Reveals Hierarchical Looping of Chromatin Fibers

Near Gene Regulatory Elements | The Journal of Physical Chemistry B [Электронный ресурс]. URL: https://pubs.acs.org/doi/10.1021/acs.jpcb.6b03197 (дата обращения: 23.05.2023).
33. Data Structures for Statistical Computing in Python [Электронный ресурс]. URL: https://www.researchgate.net/publication/340177686_Data_Structures_for_Statistical_Computing_in_Python (дата обращения: 25.05.2023).
34. Array programming with NumPy | Nature [Электронный ресурс]. URL: https://www.nature.com/articles/s41586-020-2649-2 (дата обращения: 25.05.2023).
35. Numba | Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC [Электронный ресурс]. URL: https://dl.acm.org/doi/10.1145/2833157.2833162 (дата обращения: 25.05.2023).
36. Proceedings of the Python in Science Conference (SciPy): Exploring Network Structure, Dynamics, and Function using NetworkX [Электронный ресурс]. URL: https://conference.scipy.org/proceedings/SciPy2008/paper_2/ (дата обращения: 25.05.2023).