Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный университет имени М.В. Ломоносова»

Факультет фундаментальной физико-химической инженерии

Биологический факультет МГУ имени М.В. Ломоносова

Кафедра биоинженерии

Группа интегративной биологии

БАКАЛАВРСКАЯ РАБОТА

«Анализ взаимодействий пионерных транскрипционных факторов с белками ядра клеток человека с помощью методов искусственного интеллекта»

Выполнила студентка
4 курса 401 группы
Хасанова Ума Наурузовна

_____
(подпись студента)

Научный руководитель
(д.ф.-м.н., профессор, чл.-корр. РАН)
Шайтан Алексей Константинович

_____
(подпись руководителя)

Научный консультант
Грибкова Анна Кирилловна
(м.н.с)

_____
(подпись консультанта)

Допущена к защите_____

_____

Москва
2025

Оглавление

1. Введение

Список принятых сокращений:

ТФ — транскрипционные факторы,

ПТФ — пионерные транскрипционные факторы,

ИПСК — индуцированные плюрипотентные стволовые клетки,

ТД — трансактивационный домен,

КЛВМ — короткие линейные взаимодействующие мотивы (short linear interactions motives, SLIMs)

Пионерные транскрипционные факторы (ПТФ) представляют собой уникальный подкласс транскрипционных факторов, способных специфически связываться с целевыми последовательностями ДНК в условиях гетерохроматина, что отличает их от непионерных факторов. Эта способность позволяет ПТФ повышать доступность ДНК для других белков и стабилизировать открытое состояние хроматина, играя ключевую роль в процессах клеточного репрограммирования и дифференцировки [1]. ПТФ активируют энхансеры и влияют на трехмерную укладку хроматина [2]

ПТФ инициируют деконденсацию хроматина и активацию транскрипции, однако ключевую роль в поддержании открытого состояния хроматина играют белки, рекрутируемые ПТФ, такие как ацетилтрансферазы гистонов, ремоделеры хроматина и другие транскрипционные факторы [3]. Взаимодействия ПТФ с ядерными рецепторами гормонов, например, FOXA1 с рецепторами эстрогенов и андрогенов, демонстрируют их роль в гормонозависимых процессах, включая канцерогенез [4,5]. Белки семейства KLF, такие как KLF4, могут взаимодействовать с промоторами ядерных

рецепторов, что подчеркивает их значимость в регуляции сигнальных путей [6].

К ключевым представителям пионерных транскрипционных факторов (ПТФ) относятся белки SOX2, OCT4 и KLF4. Эти белки играют важную роль в поддержании плюрипотентности, регулируя экспрессию генов NANOG и LIN28 [3]. Вместе с фактором MYC (c-MYC, который не является пионерным, поскольку взаимодействует с «закрытым» хроматином только в присутствии других факторов) они формируют «Коктейль Яманаки» — комбинацию белков, достаточную для репрограммирования дифференцированных клеток (например, фибробластов) в индуцированные плюрипотентные стволовые клетки (ИПСК) [4].

ИПСК имеют значительный потенциал для практического применения. В регенеративной медицине они позволяют выращивать ткани и органоиды с полной иммунной совместимостью, используя клетки пациента. В научных исследованиях и разработке лекарственных средств ИПСК служат ценными моделями для изучения заболеваний, особенно генетических нарушений, поскольку, в отличие от иммортализованных раковых клеточных линий, не содержат характерных для них генетических перестроек [5].

В архитектуре транскрипционных факторов (ТФ) выделяют ДНК-связывающие домены (структурно консервативные, относятся к 25 семействам) и эффекторные (активационные/репрессорные) домены, часто расположенные в неупорядоченных регионах (IDR) [7,8]. У ПТФ активационные домены называют трансактивационными (ТД). Их классифицируют на: кислотные (наиболее распространены, содержат отрицательно заряженные и гидрофобные остатки),

пролин-/серин-/глутамин-богатые (менее распространены, частично перекрываются с кислотными) [8].

ТД часто включают сайты посттрансляционных модификаций для регуляции взаимодействий [7]. Существующие модели предсказания ТД (PADDLE [9], ADPred [10]) ограничены кислотными доменами и короткими фрагментами (9–53 а.о.), тогда как медианная длина ТД — 91 а.о. [11]. Взаимодействие ТД с партнерами обеспечивается: динамичными слабыми связями, мультивалентными взаимодействиями.

Важным элементом ТД являются короткие линейные мотивы (SLIMs), которые способны активировать транскрипцию. Эти мотивы демонстрируют переход из неупорядоченного состояния в упорядоченное при связывании с партнерами, что делает их ключевыми игроками в регуляции транскрипции. Пример — 9aaTAD, встречающийся у многих ТФ (включая ПТФ): в свободном состоянии неупорядочен, но при связывании образует α-спираль [12].

Несмотря на прогресс в изучении взаимодействий ДНК-связывающих доменов ПТФ с нуклеосомами [13–15], молекулярные механизмы их взаимодействий с другими белками хроматина остаются малоизученными. Современные методы, такие как AlphaFold, позволяют предсказывать структуры белковых комплексов, что открывает новые возможности для исследования этих взаимодействий [10, 11].

Современные методы машинного обучения, такие как модели группы AlphaFold [16], позволяют предсказывать трехмерные структуры белков, включая их комплексы с модифицированными аминокислотами, нуклеиновыми кислотами и лигандами (AlphaFold 3 [17]), а также белок-белковые взаимодействия (AlphaFold-Multimer [18], AlphaFold 3 [17]).

Для оценки качества предсказаний используются метрики: plDDT (оценка достоверности на уровне аминокислотных остатков), pTM (общая топологическая точность) и ipTM (точность предсказания интерфейсов взаимодействия).

Методы AlphaFold уже доказали свою эффективность в предсказании комплексов ядерных белков, включая принципиально новые структуры. Например, с их помощью были построены интерактомы димера гистонов H2A-H2B и белков, поддерживающих целостность генома, с последующей экспериментальной проверкой [19,20].

В данной работе мы использовали алгоритм AlphaFold-Multimer для предсказания комплексов ПТФ SOX2, OCT4 и KLF4 с ядерными белками человека. Наш анализ выявил новых белков-партнеров, включая метилтрансферазы и ядерные рецепторы, а также подтвердил роль мотива 9aaTAD в формировании интерфейсов взаимодействия. Эти результаты расширяют понимание молекулярных механизмов, лежащих в основе функций ПТФ, и открывают новые направления для исследований в области клеточного репрограммирования и регуляции транскрипции.

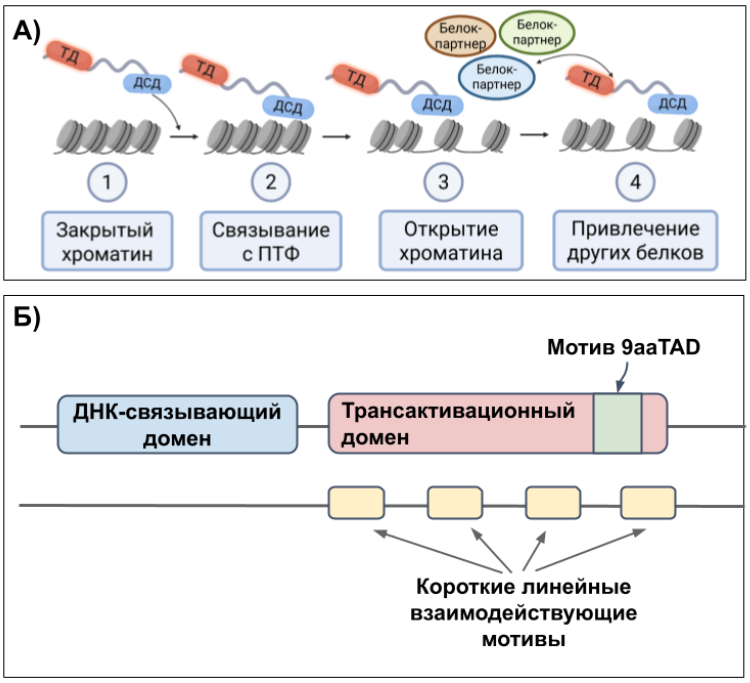

Рисунок 1. А) Схема процесса "открытия" хроматина посредством взаимодействия ПТФ с нуклеосомной ДНК. После первоначального связывания ПТФ с нуклеосомами происходит привлечение большого количества ПТФ и стабилизация взаимодействий ПТФ с хроматином. После изменения эпигенетического профиля на нуклеосомах происходит рекрутирование других белков - ремоделлеров хроматина, транскрипционных факторов и других белков. Б) Общая схема доменной организации ПТФ: ДНК-связывающий домен, трансактивационный домен (функциональный неструктурированный домен), мотив 9aaTAD. Длинный неупорядоченный регион ПТФ часто содержит короткие линейные взаимодействующие мотивы, которые способны к слабым мультивалентным взаимодействиям при разделении фаз жидкость-жидкость.

Цель данной работы: предсказание и анализ белок-белковых взаимодействий между пионерными транскрипционными факторами (SOX2, OCT4 и KLF4) и ядерными белками человека с использованием методов генеративного ИИ.

В рамках поставленной цели решались следующие **задачи:**

1. Формирование репрезентативного набора ядерных белков и предсказание комплексов ПТФ с ядерными белками;
2. Структурный анализ предсказанных белковых комплексов;
3. Функциональная характеристика и анализ представленности в экспериментальных базах данных белков-партнеров.

2. Литературный обзор

2.1. Обзор пионерных транскрипционных факторов - SOX2, OCT4, KLF4

Ключевые представители пионерных транскрипционных факторов (ПТФ) - SOX2, OCT4 (POU5F1) и KLF4 - играют центральную роль в репрограммировании клеток, инициируя ремоделирование хроматина и активируя экспрессию генов развития. Эти факторы обладают уникальной способностью связываться с нуклеосомной ДНК, воздействуя на закрытый хроматин [1,21].

В отличие от обычных транскрипционных факторов, SOX2, OCT4 и KLF4 распознают частичные или дегенерированные мотивы ДНК на поверхности нуклеосом. OCT4 связывается с нуклеосомами через свой POU-домен, вытесняя линкерный гистон и рекрутируя комплекс ремоделирования хроматина [22]. SOX2 демонстрирует гибкость в динамике связывания с ДНК, позволяющую взаимодействовать с частично экспонированными нуклеосомными мотивами [23].

В процессе репрограммирования OCT4 и SOX2 функционируют как независимо, так и кооперативно для поддержания открытых состояний хроматина. OCT4 критически важен для сохранения хроматиновой доступности при клеточном делении [24]. Участки открытого хроматина, поддерживаемые SOX2, коррелируют с активацией экспрессии генов [25]. KLF4 способен стабилизировать связывание OCT4/SOX2 с энхансерами [26].

Триада SOX2/OCT4/KLF4 наглядно демонстрирует, как пионерные транскрипционные факторы интегрируют связывание с нуклеосомами,

ремоделирование хроматина и реорганизацию трехмерной структуры генома для управления переходами между клеточными состояниями.

2.2. Пионерные транскрипционные факторы формируют эпигенетический ландшафт в клетке

Пионерные факторы играют центральную роль в формировании эпигенетической памяти - устойчивых изменений в структуре хроматина, которые сохраняются после прекращения действия самого фактора и передаются при клеточных делениях [2]. Этот процесс осуществляется через сложный каскад молекулярных событий. На начальном этапе пионерные факторы, такие как FOXA1/2, PAX7 или ASCL1, связываются с компактизированными участками хроматина, содержащими факультативный гетерохроматин. Важно отметить, что они избегают участков конститутивного гетерохроматина, который представляет непреодолимый барьер. После связывания пионерные факторы инициируют процесс ремоделирования хроматина, который включает несколько ключевых этапов. Во-первых, происходит локальное ослабление взаимодействий между ДНК и гистонами, что было продемонстрировано на примере FOXA1, который может вытеснять ДНК из нуклеосомы. Во-вторых, запускается активное или пассивное деметилирование ДНК. В случае активного механизма пионерные факторы рекрутируют ферменты семейства TET (TET1-3), которые последовательно окисляют 5-метилцитозин до 5-гидроксиметилцитозина и далее до неметилированного состояния. Например, FOXA1 непосредственно взаимодействует с TET1, направляя его к специфическим энхансерам в клетках печени. Альтернативный пассивный механизм предполагает ингибирование ДНК-метилтрансферазы DNMT1 и ее кофактора UHRF1, что

предотвращает восстановление метильных меток после репликации. Параллельно с изменениями в метилировании ДНК происходят характерные модификации гистонов. На ранних стадиях появляется слабый моносигнал H3K4me1, опосредованный гистон-метилтрансферазными комплексами MLL3/4, что соответствует состоянию "приммированного" энхансера. По мере стабилизации открытого состояния хроматина формируется бимодальное распределение H3K4me1 с участком нуклеосомного вытеснения в центре энхансера, что является признаком его полной активации. Ключевым событием становится рекрутирование коактиваторов p300/CBP, которые ацетилируют H3K27, создавая маркер H3K27ac активного энхансера.

Интересно, что даже после прекращения действия пионерного фактора или сигнального стимула, многие из этих изменений сохраняются, формируя эпигенетическую память [2]. Например, в клетках печени деметилированные FOXA1-зависимые энхансеры остаются доступными даже при нокдауне FOXA1. В иммунных клетках "приммированные" энхансеры (с H3K4me1, но без H3K27ac) сохраняют память о предыдущей активации и обеспечивают более быстрый ответ при повторном воздействии антигена.

Важным аспектом является наследование этих эпигенетических меток при клеточных делениях. Деметилированная ДНК и определенные модификации гистонов (например, H3K4me1) могут передаваться через репликативную вилку благодаря механизмам эпигенетического копирования. Некоторые пионерные факторы, такие как SOX2 и OCT4, обладают дополнительной функцией митотического букмаркинга [27,28] - они остаются связанными с хроматином во время митоза, что облегчает восстановление транскрипционной программы после деления клетки.

Таким образом, пионерные факторы не просто временно активируют гены, а создают устойчивые эпигенетические изменения, которые определяют долговременную клеточную идентичность и лежат в основе таких процессов, как клеточная дифференцировка, пластичность и репрограммирование. Нарушения этих механизмов могут приводить к серьезным патологиям, включая рак и нейродегенеративные заболевания, что подчеркивает их фундаментальное значение в биологии развития и медицине [2].

2.3. Предсказание структур белковых комплексов с помощью AlphaFold

Современные методы структурной биологии, включая крио-ЭМ и AlphaFold, открывают новые возможности для изучения механизмов действия пионерных транскрипционных факторов (ПТФ). В частности, последняя версия AlphaFold 3 (AF3) [17, p. 3] демонстрирует значительный прогресс в предсказании структур биомолекулярных комплексов, преодолевая ограничения традиционных вычислительных методов, таких как молекулярный докинг (AutoDock Vina) и статистические модели, которые зависели от существующих структурных данных и имели ограниченную точность [18,29].

AF3, основанная на диффузионной архитектуре и использующая модуль "pairformer", существенно снижает зависимость от множественных выравниваний последовательностей (MSA) и обеспечивает более точное предсказание взаимодействий [17]. Эта модель превосходит специализированные инструменты, демонстрируя высокую точность в предсказании комплексов антиген-антитело, белок-ДНК, а также

взаимодействий с нуклеиновыми кислотами, лигандами и модифицированными остатками.

Важным достижением является применение AlphaFold-Multimer для масштабного предсказания структур комплексов ядерных белков. В исследовании Burke et al., 2023 [30] было успешно предсказано 65,484 уникальных пар белковых взаимодействий с использованием метрики pDockQ (порог >0.5 означает 80% точность). Это открывает новые возможности для изучения взаимодействий ПТФ (SOX2, OCT4, KLF4) с другими ядерными белками, что ранее не проводилось систематически.

Однако остаются нерешенные вопросы:

1. Склонность к "галлюцинациям" в неупорядоченных областях, характерных для многих ПТФ;
2. Ограничения в предсказании крупных мультибелковых комплексов;
3. Трудности определения стехиометрии взаимодействий.

Для преодоления этих ограничений требуется интеграция с экспериментальными методами (например, масс-спектрометрией сшивок для валидации интерфейсов) и другими вычислительными подходами [30]. Тем не менее, AlphaFold2, AlphaFold3 представляют собой мощные инструменты для изучения белок-белковых взаимодействий.

3. Материалы и методы

3.1. Формирование репрезентативного набора ядерных белков и предсказание комплексов ПТФ с ядерными белками

3.1.1. Разметка координат функциональных и структурных доменов белков ПТФ (SOX2, OCT4 и KLF4) человека.

Для выбранных пионерных транскрипционных факторов границы структурных и функциональных доменов определялись путем объединения аннотированной информации из баз данных Pfam [31], InterPro [32] и TFRegDB [11]. На основании этих данных были установлены положения ДНК-связывающих и трансактивационных доменов, представленные в **Таблице 1**. Кроме того, в аминокислотных последовательностях исследуемых белков были отмечены участки, соответствующие мотиву 9aaTAD.

Таблица 1. Доменная организация OCT4, SOX2, и KLF4.

Белок, ген	**Координаты ДНК-связывающего домена (по Pfam)**	**Координаты ТД (по TFRegDB)**	**Координаты мотива 9aaTAD**
OCT4 (POU5F1)	143-212 (POU-specific, POU_S), 231-287 (гомеодомен, POU_{HD})	1-138, 290-360	4-12
SOX2 (SOX2)	41-109 (HMG box)	118-317	272-280

Белок, ген	Координаты ДНК-связывающего домена (по Pfam)	Координаты ТД (по TFRegDB)	Координаты мотива 9aaTAD
KLF4 (KLF4)	430-454, 460-484, 490-512 (цинковый палец, тип C2H2)	92-112	101-109

3.1.2. Построение репрезентативной выборки ядерных белков на основе UniProt, Human Protein Atlas и классификации SimChrom

Для выполнения задачи был сформирован список потенциальных белков-партнёров для пионерных транскрипционных факторов. Отбор проводился на основе аннотаций из баз данных UniProt [33] и Human Protein Atlas [34], а также с использованием классификации белков хроматина SimChrom (см. раздел Функциональная характеристика белков-партнеров). В выборку включались только белки, локализованные в ядре, за исключением белков, принадлежащих к ядерной мембране, ядрышку и ядерным тельцам. Для снижения вычислительной нагрузки учитывались лишь те белки, длина которых не превышала 1530 аминокислот, что соответствует 95 процентилю распределения длины последовательностей. Далее для устранения избыточности последовательности были сгруппированы по 90-процентной идентичности с помощью алгоритма кластеризации CD-HIT [35]. Кроме того, были исключены белки, для которых модель RoseTTAFold2-PPI [36] предсказывала низкокачественные взаимодействия. В результате отбора был

получен финальный набор из 3557 ядерных белков человека, использованный для последующего моделирования комплексов с ПТФ.

3.1.3. Предсказание комплексов ПТФ с ядерными белками

Структуры комплексов между отобранными пионерными транскрипционными факторами и ядерными белками предсказывались с использованием модели AlphaFold Multimer v3 [18]. Для выполнения расчетов применялся локальный запуск через интерфейс ColabFold версии 1.5.5 с использованием графических ускорителей класса Nvidia RTX A5000. Поиск гомологов осуществлялся с помощью локального сервера MMseqs [37]. В качестве входных данных использовались аминокислотные последовательности белков в формате FASTA. Предсказания выполнялись с тремя циклами рекурсии и с применением структурных шаблонов из базы PDB. Оценка качества полученных моделей проводилась на основе трех метрик: ipTM, pDockQ и ipSAE, для которых были заданы соответствующие пороговые значения.

Метрика ipTM используется в модели AlphaFold Multimer для оценки точности топологии взаимодействующего интерфейса в предсказанном белковом комплексе. Значение pDockQ представляет собой аппроксимированную версию показателя DockQ, полученную с помощью сигмоидальной функции. Метрика DockQ является комплексной оценкой качества докинга, объединяющей три основные характеристики, используемые в рамках инициативы CAPRI (Critical Assessment of Predicted Interactions, критическая оценка предсказанных взаимодействий): fnat (fraction of native contacts) — доля нативных контактов между цепями, присутствующих в предсказанной структуре; LRMSD (Ligand Root Mean

Square Deviation, среднеквадратичное отклонение лиганда) — среднеквадратичное отклонение положения лиганда после наложения рецептора; iRMSD (interface Root Mean Square Deviation, среднеквадратичное отклонение интерфейса взаимодействия) — среднеквадратичное отклонение атомов интерфейса между моделью и референсной структурой.

Показатель ipSAE (interface predicted TM Score based on Aligned Errors) был разработан как усовершенствованная версия ipTM и предназначен для более надежной оценки качества взаимодействий в случаях наличия неструктурированных участков или дополнительных доменов, не вовлеченных во взаимодействие. В отличие от ipTM, где расчет масштабирующего коэффициента d_0 зависит от полной длины белковых цепей, в ipSAE d_0 вычисляется только на основе остатков с приемлемыми значениями предсказанных выровненных ошибок (PAE, predicted aligned error). Это позволяет минимизировать влияние нерелевантных участков на итоговое значение. В данной работе при расчёте ipSAE использовались значения pae_cutoff = 15 Å, dist_cutoff = 10 Å. Метрика определялась в обоих направлениях взаимодействия (A→B и B→A), после чего выбиралось максимальное значение для повышения чувствительности метода к истинным контактам.

На основе литературных данных [38,39] были выбраны следующие пороговые значения для качества комплексов: комплексы среднего качества отбирали при выполнении хотя бы одного условия pDockQ > 0,23, ipSAE > 0,3 или ipTM > 0,6; комплексы высокого качества — pDockQ > 0,5, ipSAE > 0,5 или ipTM > 0,8.

3.1.4. Валидация достоверности предсказанных комплексов

Был составлен список белков, не взаимодействующих с выбранными ПТФ, с целью валидации предсказаний AlphaFold. Для исследования был составлен набор контрольных белков с экспериментально подтвержденной локализацией в пяти ключевых компартментах (митохондрии, промежуточные филаменты, аппарат Гольджи, везикулы) по данным Human Protein Atlas [34]. Использовались только белки с аннотированными UniProt ID, для которых отсутствовали известные взаимодействия с Sox2, Oct4 и Klf4 в базах данных STRING [40] и BioGRID [41]. Контрольные белки подбирались таким образом, чтобы минимизировать вероятность ложноположительных результатов при предсказании взаимодействий. Для отобранных белков были построены структуры комплексов с ПТФ (SOX2, OCT4, KLF4). Была произведена оценка качества полученных комплексов с помощью метрик качества pDockQ, ipTM, ipSAE с пороговыми значениями, аналогичными вышеописанному этапу исследования.

3.2. Структурный анализ белковых комплексов

3.2.1. Анализ распределения контактов вдоль последовательности ПТФ

Для количественной оценки взаимодействий в предсказанных комплексах между пионерными транскрипционными факторами и ядерными белками был выполнен анализ межатомных контактов. Контакт устанавливался, если пара атомов, принадлежащих разным белкам, находилась на расстоянии ≤ 4 Å. При этом белки считались взаимодействующими, если хотя бы один аминокислотный остаток одного из

них образовывал контакт с партнерским белком. Общее число контактов для каждого остатка ПТФ суммировалось по всем белкам-партнёрам, после чего полученные значения нормированы на общее количество партнёрских белков.

3.2.2. Анализ аминокислотного состава интерфейсов взаимодействия

Для исследования аминокислотного состава взаимодействующих были отобраны участки белков-партнеров, взаимодействующие с ДНК-связывающим доменом, трансактивационным доменом и мотивом 9aaTAD. Отобранные участки были проанализированы на обогащение различными типами аминокислотных остатков: гидрофобными (G, A, V, L, I, M, F, W, P), отрицательно заряженными (D, E), полярными (S, T, N, Q, Y, C), положительно заряженными (K, R, H). Таким образом, был проанализирован аминокислотный состав взаимодействующих остатков партнеров для каждого домена.

3.2.3. Исследование динамики неупорядоченных регионов при образовании комплексов

С учетом высокой доли внутренне неупорядоченных регионов в структуре исследуемых пионерных транскрипционных факторов была сформулирована гипотеза о том, что отдельные участки этих белков могут переходить из неупорядоченного в упорядоченное состояние при взаимодействии с белками-партнерами. Для проверки этого предположения был реализован подход, основанный на сопоставлении структурных

характеристик ПТФ в свободном состоянии и в составе предсказанных белковых комплексов.

В качестве основного критерия для выявления неупорядоченных участков использовался показатель plDDT. Остатки с plDDT ниже 50 в изолированной структуре классифицировались как неупорядоченные. Если при формировании комплекса значение plDDT превышало этот порог, фиксировался переход в упорядоченное состояние. Для обеспечения сопоставимости результатов данные о таких переходах нормировались на общее число проанализированных моделей.

Дополнительно была выполнена проверка совпадений участков, претерпевающих конформационные изменения, с короткими линейными взаимодействующими мотивами, информация о которых была получена из базы MobiDB [42].

При анализе учитывались как пространственное расположение переходящих участков относительно функциональных доменов ПТФ, так и их совпадения с известными мотивами межбелкового взаимодействия. Наибольшее внимание уделялось тем регионам, где наблюдалась высокая частота переходов, поскольку они могут быть важны для понимания механизмов регуляции и динамики взаимодействий белков.

3.2.4. Энергетическая оптимизация структурных моделей с использованием FoldX и анализ вклада различных типов взаимодействий в стабилизацию комплексов

Мы провели энергетическую оптимизацию комплексов пионерных транскрипционных факторов (ПТФ) с белками хроматина. Из всех предсказанных моделей были отобраны 40 комплексов, удовлетворяющих как

минимум двум из трех критериев качества: высокие значения ipTM, DockQ и pDockQ. Для них была проведена структурная оптимизация в силовом поле FoldX v5.1 [43], позволяющая устранить локальные стерические конфликты и привести геометрию комплекса к энергетически благоприятному состоянию.

3.2.5. Оценка влияния онкогенной мутации на взаимодействия KLF4

Для исследованных ПТФ известна только одна точечная рекуррентная онкологическая мутация: в KLF4 замена K409Q/N, обнаруженная в образцах рака молочной железы, уротелиальной карциномы мочевого пузыря, аденокарциномы предстательной железы (по данным cBioPortal) [44]. Мы проанализировали белок-белковые взаимодействий этого сайта с другими ядерными белками. В ходе исследования с использованием программы FoldX v5.1 была проведена оценка энергетического влияния точечных мутаций K443N и K443Q в белке Klf4 на его взаимодействие с 26 белками-партнерами. Анализ проводился на основе изменения свободной энергии связывания (ΔΔG), структурных параметров (RMSD, BSA) и вклада различных энергетических компонентов.

3.3. Функциональная характеристика белков-партнеров

3.3.1. GO-обогащение и функциональная классификация интеракторов ПТФ

Для проведения функциональной аннотации белков-партнёров пионерных транскрипционных факторов, предсказанных в составе белковых комплексов, был использован анализ обогащения по терминам Генной

онтологии в аспектах «молекулярная функция» и «биологический процесс». Для этого применялся инструмент enrichr из библиотеки GSEAPY [45]. Значимыми считались только те термины, которые имели с adjusted p-value < 0,05. Для классификации белков хроматина дополнительно использовался специализированный набор, сформированный полуавтоматическим способом на основе аннотаций Gene Ontology, профильных баз данных и публикаций. В работе применялась классификация SimChrom, включающая 39 функциональных групп и 3045 белков хроматина. Актуальная версия классификации доступна по адресу https://simchrom.intbio.org/#classification (по состоянию на 14 апреля 2025 года).

3.3.2. Анализ представленности полученных комплексов в экспериментальных базах данных

С целью оценки наличия предсказанных взаимодействий в экспериментально подтвержденных источниках была выполнено сравнение комплексов среднего и высокого качества с данными, содержащимися в базах STRING [40] и BioGRID [41]. При этом из базы STRING учитывались только взаимодействия с уровнем достоверности не ниже 0.9, тогда как из BioGRID включались все физические взаимодействия, независимо от метода их обнаружения.

4. Результаты и обсуждение

4.1. Анализ предсказанных структурных комплексов пионерных транскрипционных факторов человека (SOX2, OCT4, KLF4) с ядерными белками

С помощью молекулярного моделирования и структурного анализа были изучены белок-белковые взаимодействия пионерных транскрипционных факторов (ПТФ) с белками хроматина с целью выявления возможных молекулярных механизмов, запускаемых после связывания ПТФ с нуклеосомной ДНК в составе гетерохроматина. Общая схема исследования приведена на Рисунке 2. В фокусе исследования находились три ПТФ из так называемого «Коктейля Яманаки» — SOX2, OCT4 и KLF4, играющие ключевую роль в поддержании плюрипотентного состояния и индуцированном репрограммировании клеток. Для предсказания взаимодействий пионерных транскрипционных факторов (ПТФ) человека с другими ядерными белками был проведён отбор 3557 белков с подтвержденной ядерной локализацией на основе данных из баз UniProt и Human Protein Atlas. Предсказание структур белок-белковых комплексов осуществлялось с использованием модели AlphaFold 2 Multimer (см. «Материалы и методы»).

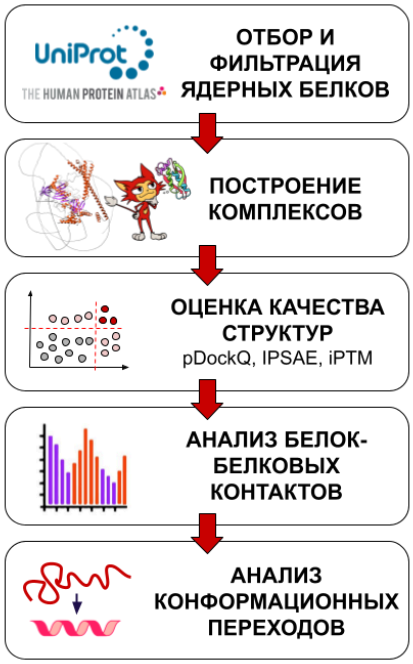

Рисунок 2. Общая схема проведенного исследования

Учитывая, что по данным литературы белок-белковые взаимодействия пионерных транскрипционных факторов (ПТФ) часто опосредуются их неупорядоченными регионами, для оценки достоверности предсказанных структурных моделей мы применили три метрики: ipTM, pDockQ и ipSAE (см. раздел «Материалы и методы» для подробного описания). На основании этих метрик все 10 632 предсказанных комплекса были классифицированы по качеству: низкое (9179 структур), среднее и высокое (1453 структуры), из которых 267 структур показали высокие значения всех трех метрик (Рисунок 3). Структуры с низким качеством по всем метрикам были исключены из последующего анализа. Оставшиеся комплексы среднего и высокого качества включали 487 структур для SOX2, 640 — для OCT4 и 326 — для KLF4. Пересечение белков-партнёров между тремя ПТФ представлено на Рисунке 4A. Как видно, общее множество партнеров преимущественно участвует в неспецифических взаимодействиях с трансактивационным доменом (TA) или

мотивом 9aaTAD, что обсуждается далее. В то же время, уникальные взаимодействия каждого из ПТФ с определенными белками обусловлены различиями в их ДНК-связывающих доменах (см. Таблицу 1 и пояснения ниже).

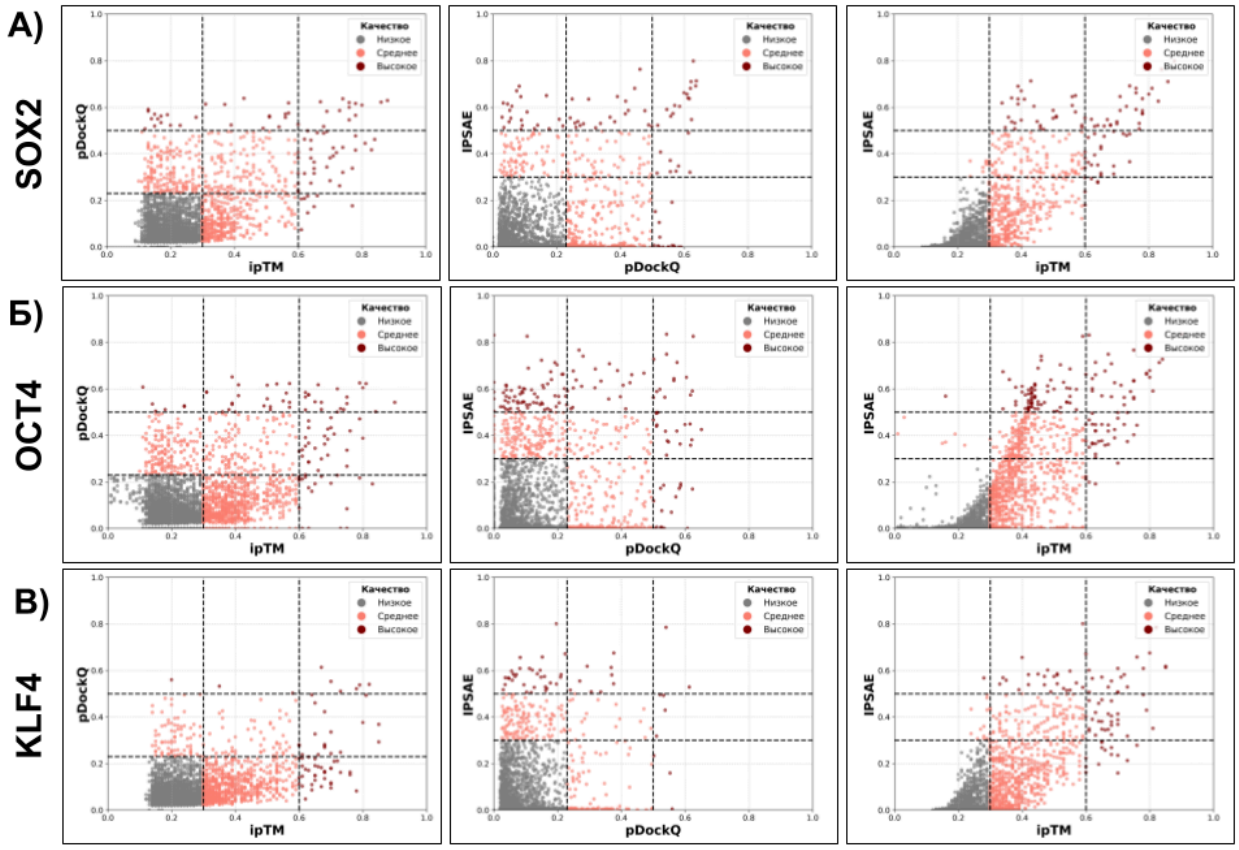

Рисунок 3. Метрики качества (ipTM, pDockQ и ipSAE) предсказанных структур для ПТФ (SOX2, KLF4, OCT4).

Структурные комплексы ПТФ с ядерными белками, получившие наивысшие оценки по всем трём метрикам качества, доступны в интерактивной форме в дополнительных материалах (https://intbio.org/Khasanova_et_al_2025). Краткий перечень отобранных

моделей с соответствующими значениями метрик приведён в Таблице 2, а отдельные примеры визуализированы на Рисунке 4Б. Среди структур с наибольшими значениями ipTM можно выделить комплексы с метилтрансферазами факторов элонгации: SOX2-EEF1AKMT3 (ipTM = 0,86) и EEF2KMT — OCT4-EEF2KMT (ipTM = 0,90), SOX2-EEF2KMT (ipTM = 0,81). Во взаимодействии SOX2 с EEF2KMT участвуют HMG-домен и мотив 9aaTAD, тогда как OCT4 взаимодействует через оба своих ДНК-связывающих домена; при этом контактирующая поверхность на EEF2KMT оказывается одинаковой в обоих случаях. Также высокие метрики качества показали комплексы ПТФ с гистон-метилтрансферазами: SOX2-N6AMT1 (ipTM = 0,86), где N6AMT1 метилирует гистон H4 по K12 — метка, ассоциированная с активными промоторами; и SOX2-METTL23 (ipTM = 0,83), METTL23 модифицирует гистон H3 по R17, способствуя активации транскрипции. Эти взаимодействия ранее не описывались в литературе и требуют дальнейшего изучения для выяснения их возможной функциональной роли.

Таблица 2. Предсказанные комплексы ПТФ с ядерными белками человека, отобранные по метрикам качества.

ПТФ	**Ген партнера, Uniprot AC**	**Функция белка-партнера**	**ipTM**	**ipSAE**	**pDockQ**
OCT4	EEF2KMT, Q96G04	метилтрансфераза фактора элонгации EEF2	0,9	0,83	0,5

OCT4	PPP4C, P60510	фосфатаза белков	0,84	0,73	0,5
OCT4	VCPKMT, Q9H867	метилтрансфераза белков	0,83	0,7	0,19
OCT4	RPS16, P62249	Белок малой рибосомальной субъединицы uS9	0,81	0,6	0,6
KLF4	CDKN2D, P55273	Ингибитор циклин-зависимой киназы 4 D	0,85	0,61	0,29
KLF4	MAD2L2, Q9UI95	Белок контрольной точки сборки митотического веретена MAD2B	0,85	0,61	0,36
KLF4	SLX1A/SLX1B, Q9BQ83	Структурно-специфическая субъединица эндонуклеазы SLX1	0,82	0,78	0,54
SOX2	NUP37, Q8NFH4	Нуклеопорин	0,88	0,8	0,63
SOX2	N6AMT1, Q9Y5N5	Метилтрансфераза белков, в том числе	0,86	0,71	0,62

		гистона H4K12 (метка промоторов)			
SOX2	EEF1AKMT3, Q96AZ1	метилтрансфераза факторов элонгации EEF1A1	0,84	0,76	0,46
SOX2	METTL23, Q86XA0	Метилтрансфераза гистонов H3R17	0,83	0,63	0,42
SOX2	EEF2KMT, Q96G04	метилтрансфераза факторов элонгации EEF2	0,81	0,68	0,45

В число комплексов с высокими метриками качества вошел и комплекс KLF4-CDKN2D, в котором наблюдаются признаки артефактов моделирования: область взаимодействия включает первый цинковый палец KLF4, расположенный почти в одной плоскости, а также множественные наложения атомов. Подобные искажения, вероятно, могли бы быть устранены с помощью процедуры релаксации предсказанных структур, однако применение релаксации ко всем моделям потребовало бы крайне значительных вычислительных ресурсов.

Для валидации метода мы собрали контрольный набор из 40 белков с подтвержденной локализацией в митохондриях, филаментах, аппарате Гольджи и везикулах, не имеющих известных взаимодействий с исследуемыми ПТФ по базам STRING и BioGRID.

Мы проанализировали отобранные комплексы из контрольного набора по ключевым метрикам: pDockQ, ipTM и ipSAE. Были рассчитаны основные статистические показатели для каждой метрики. Для pDockQ среднее значение составило 0.154, медиана 0.120, дисперсия 0.016. Минимальное значение 0.024, максимальное 0.502. Для ipTM среднее 0.22, медиана 0.20, дисперсия 0.019. Диапазон от 0.09 до 0.70. Наименее стабильной оказалась метрика ipSAE со средним 0.038, медианой 0.005 и высокой дисперсией 0.008, варьируя от 0 до 0.476.

Применяя пороговые значения для определения комплексов среднего качества (pDockQ > 0.23, ipTM > 0.6, ipSAE >0.3), выявлено следующее. По pDockQ критерию удовлетворяют 14 комплексов (35% от общего числа), включая O43474_P15291 (0.257), Q01860_P04179 (0.502) и Q01860_P14136 (0.428). По ipTM только один комплекс превысил порог 0.6 - Q01860_P04179 (0.70). Для ipSAE три комплекса показали значения выше 0.3: O43474_P15291 (0.476), Q01860_P04179 (0.457) и Q01860_P05783 (0.089). Особого внимания заслуживает комплекс Q01860_P04179, который является единственным, превысившим порог по ipTM (0.70), а также показавший выдающиеся значения по pDockQ (0.502) и ipSAE (0.457). Комплекс O43474_P15291 также демонстрирует хорошие характеристики с pDockQ 0.257 и ipSAE 0.476, хотя его ipTM (0.58) немного не дотягивает до порогового значения. Анализ корреляции между метриками показывает, что pDockQ и ipSAE имеют умеренную положительную корреляцию (визуально наблюдаемое увеличение pDockQ при росте ipSAE), в то время как связь между ipTM и pTM менее выражена. Стоит отметить, что большинство комплексов (32 из 40) имеют ipSAE ниже 0.01, что указывает на общую слабость взаимодействий в этой метрике.

В рамках анализа структурных моделей комплексов, отобранных по средним и высоким метрикам качества, мы исследовали, какие участки ПТФ участвуют во взаимодействии с ядерными белками (см. Рисунок 4В–Д). Для всех трёх факторов — SOX2, OCT4 и KLF4 — было показано, что основную роль в формировании интерфейса взаимодействия играет ДНК-связывающий домен. В частности, в 96% комплексов KLF4, 94% комплексов SOX2 и 93% комплексов OCT4 контактирующие аминокислотные остатки локализованы в пределах ДНК-связывающего домена. Подобная картина соответствует существующим экспериментальным данным: например, ранее было показано, что OCT4 и SOX2 способны образовывать взаимодействия друг с другом посредством своих ДНК-связывающих доменов [46], что подтверждает правдоподобность полученных предсказаний.

Мы также оценили участие трансактивационного домена (ТД) и мотива 9aaTAD в белок-белковых взаимодействиях предсказанных комплексов (см. Рисунок 3В–Д). Анализ показал, что ТД вовлечен в формирование интерфейса в 56% случаев для KLF4, 62% — для SOX2 и 58% — для OCT4. Однако характер участия 9aaTAD в этих взаимодействиях существенно различается между ПТФ: у KLF4 он задействован в 53% комплексов, у SOX2 — в 45%, тогда как у OCT4 — лишь в 9%. Таким образом, в случае с KLF4 можно заключить, что основная масса взаимодействий через трансактивационный домен приходится именно на мотив 9aaTAD, что подчеркивается четко выраженным пиком контактов в этой области (Рисунок 4Д). Напротив, у OCT4 контакты с партнёрами распределены более равномерно по всей длине ТД, не формируя отчетливых локализаций, что может указывать на вспомогательную, а не ведущую роль этого региона в белок-белковых взаимодействиях (Рисунок 4Г). Сравнительно низкая частота

взаимодействий OCT4 через 9aaTAD, вероятно, обусловлена несколькими факторами: во-первых, вариациями аминокислотной последовательности мотива, что, как известно, снижает его способность к активации транскрипции [12]; во-вторых, наличием двух ДНК-связывающих доменов, которые, по-видимому, обеспечивают устойчивые контакты с партнёрами, снижая потребность в участии трансактивационного мотива.

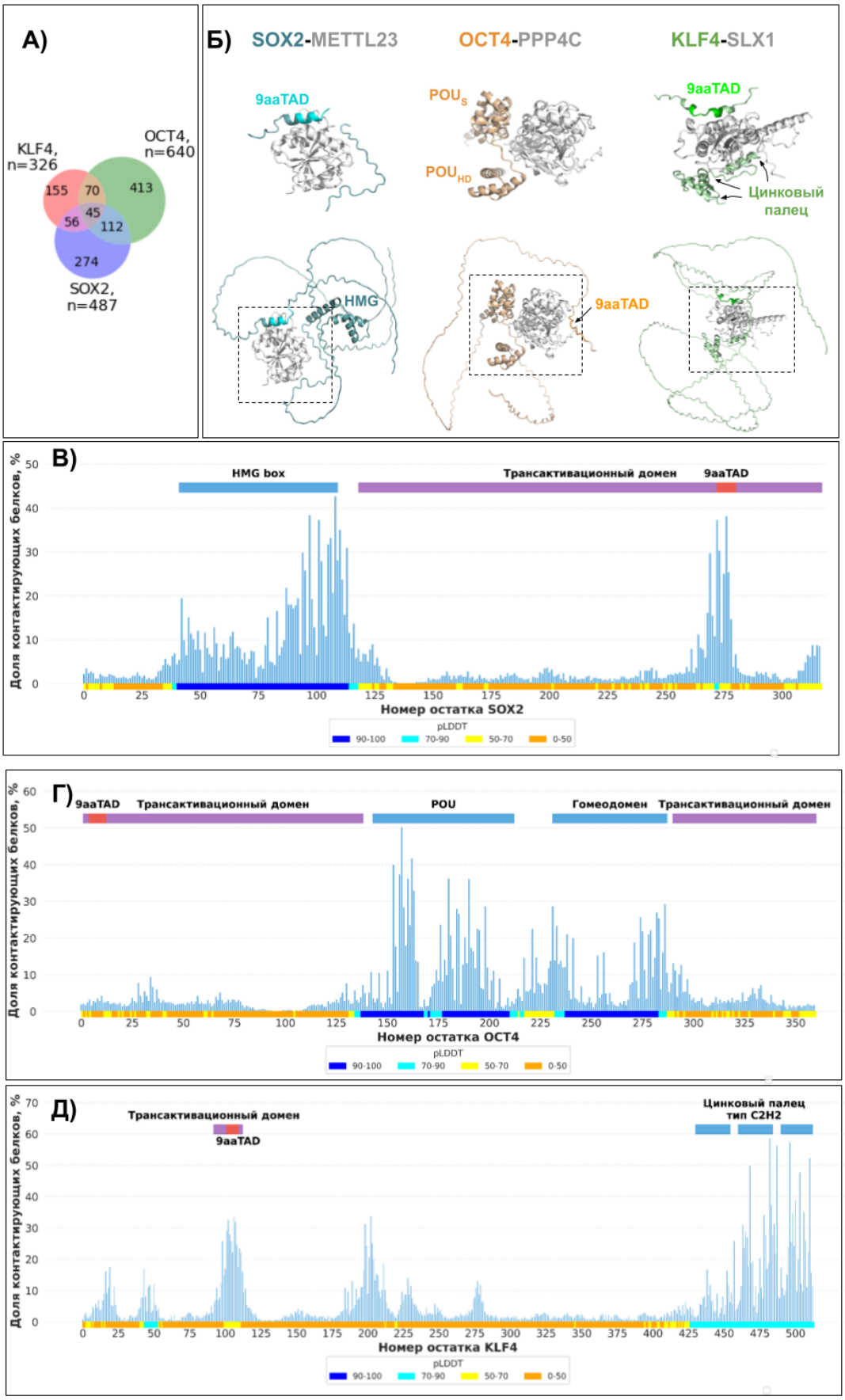
А)
KLF4, n=326
OCT4, n=640
SOX2, n=487
155
70
413
45
56
112
274
Б)
SOX2-METTL23
OCT4-PPP4C
KLF4-SLX1
9aaTAD
POU$_S$
POU$_{HD}$
9aaTAD
Цинковый палец
HMG
9aaTAD
В)
HMG box
Трансактивационный домен
9aaTAD
Доля контактирующих белков, %
Номер остатка SOX2
pLDDT
90-100
70-90
50-70
0-50
Г)
9aaTAD
Трансактивационный домен
POU
Гомеодомен
Трансактивационный домен
Доля контактирующих белков, %
Номер остатка OCT4
pLDDT
90-100
70-90
50-70
0-50
Д)
Трансактивационный домен
9aaTAD
Цинковый палец тип C2H2
Доля контактирующих белков, %
Номер остатка KLF4
pLDDT
90-100
70-90
50-70
0-50

Рисунок 4. А) Диаграмма Венна, отражающая пересечение белков-партнёров пионерных транскрипционных факторов (ПТФ) в предсказанных комплексах, отобранных по средним и высоким метрикам качества. **Б)** Примеры структур предсказанных комплексов ПТФ с другими белками, получивших высокие оценки качества. В верхней части изображена увеличенная область взаимодействия, в нижней — полная структура комплекса, включая протяженные неструктурированные участки; область контакта выделена рамкой. Отмечены ДНК-связывающие домены ПТФ и мотив 9aaTAD. **В, Г, Д)** Графики распределения доли взаимодействующих белков-партнёров вдоль последовательностей ПТФ: SOX2 (В), OCT4 (Г) и KLF4 (Д). На графиках указана доля белков, образующих контакты с ПТФ на каждом аминокислотном остатке. Под графиками представлена шкала pLDDT, характеризующая достоверность предсказания пространственной структуры изолированного ПТФ моделью AlphaFold; значения pLDDT < 50 соответствуют неупорядоченным регионам. Над графиками показана доменная организация соответствующего белка.

Детальный анализ интерфейсов взаимодействия выявил специфические паттерны аминокислотного состава. Для OCT4 взаимодействующие с ДНК-связывающим доменом остатки белков-партнеров имеют следующее распределение по типам: гидрофобные - 37.5%, отрицательно заряженные - 22.9%, полярные - 26.4%, положительно заряженные - 13.2%; для SOX2: 37.7%, 15.6%, 29.7%, 16.9%; для KLF4: 31%, 21.2%, 28%, 19.8%. Взаимодействующие с трансактивационным доменом остатки белков-партнеров имеют следующее распределение по типам: в OCT4:

гидрофобные - 36.3%, отрицательно заряженные - 15.3%, полярные - 29.2%, положительно заряженные - 19.2%; в SOX2: 42.4%, 12%, 28.7%, 16.9%; в KLF4: 40.6%, 8%, 24.2%, 27.2%. Для взаимодействующим с мотивом 9aaTAD остатков было обнаружено следующее распределение по типам: в OCT4: гидрофобные - 42.4%, отрицательно заряженные - 9.9%, полярные - 36.6%, положительно заряженные - 11%; в SOX2: 47.6%, 5.6%, 26.8%, 20%; в KLF4: 42.9%, 6.3%, 24.5%, 26.2%

Мы выполнили качественную классификацию белков-партнёров ПТФ, взаимодействующих через мотив 9aaTAD, по их предполагаемым функциям. Для этого использовались категории набора белков хроматина SimChrom, сформированного полуавтоматически на основе аннотаций Gene Ontology (см. раздел «Материалы и методы»).

Среди партнёров SOX2, взаимодействующих с его мотивом 9aaTAD, значительную долю составляют транскрипционные факторы (14%) и белки, локализующиеся в области центромер (6%). В случае OCT4 наиболее выражена доля транскрипционных факторов (21%), причём 16% составляют факторы, не относящиеся к генам домашнего хозяйства. Аналогичная тенденция наблюдается и у KLF4: 15% белков-партнёров — это транскрипционные факторы, из которых 10% — вне группы генов домашнего хозяйства; кроме того, выделяются ядерные гормональные рецепторы (5%), белки прерибосомы (5%) и белки, ассоциированные с центромерой (5%).

Предсказанное количество взаимодействующих транскрипционных факторов варьируется: для SOX2 — 32, для KLF4 — 27, для OCT4 — 12. При этом каждый из факторов связывается с уникальным набором ТФ, и лишь один белок — NTN3 — оказался общим партнером для SOX2 и OCT4.

Мы также провели энергетическую оптимизацию структурных моделей 40 высококачественных комплексов с использованием FoldX и анализ вклада различных типов взаимодействий в стабилизацию комплексов.

Анализ логов FoldX показал, что основной вклад в стабилизацию внесли улучшение упаковки и перераспределение ван-дер-ваальсовых и электростатических взаимодействий. Последовательная оптимизация остатков привела к энергетической стабилизации, отраженной в изменениях значений таких параметров, как общая энергия, энергия водородных связей, энтропийные поправки и сольватационные вклады.

4.2. Взаимодействия пионерных транскрипционных факторов (ПТФ) с ядерными белками через короткие линейные мотивы

Разработанный протокол идентификации переходов из неупорядоченного в упорядоченное состояние позволил обнаружить как упорядоченные, так и неупорядоченные сегменты в белковой структуре ПТФ, а также зафиксировать случаи перехода отдельных фрагментов из неупорядоченного состояния в упорядоченное при формировании комплексов с другими белками. Согласно нашим данным, в исследуемых ПТФ — SOX2, OCT4 и KLF4 — найдено 16 коротких участков (от 4 до 18 аминокислот), которые приобретают упорядоченную структуру при взаимодействии с партнерами. Восемь из этих фрагментов, согласно базе данных MobiDB, соответствуют известным коротким линейным мотивам (КЛВМ), служащим площадками для высокоспецифичных взаимодействий. Важно подчеркнуть, что в процессе идентификации этих конформационных изменений не принималась во внимание изначальная упорядоченность крупных доменов, таких как мотив 9aaTAD и ДНК-связывающий домен. Однако

дополнительный анализ выявил неожиданный факт: около 1–4% структурных ансамблей демонстрировали возникновение упорядочивания непосредственно вблизи границ мотива 9aaTAD в белке KLF4. Данный феномен наводит на мысль, что жесткая структура 9aaTAD сама выступает своеобразным "якорем", оказывая влияние на организацию близлежащих, первоначально неупорядоченных участков полипептида. Этот вывод подкрепляется дополнительными наблюдениями: связывание мотива 9aaTAD с белками-партнерами сопровождается индукцией структурных перестроек в окружающих областях, постепенно приводя их к упорядоченному состоянию. Таким образом, помимо собственно мотивации 9aaTAD, он косвенно способствует повышению устойчивости всей молекулы KLF4, выступая своего рода организационным центром для координации конформационного поведения прилежащих элементов.

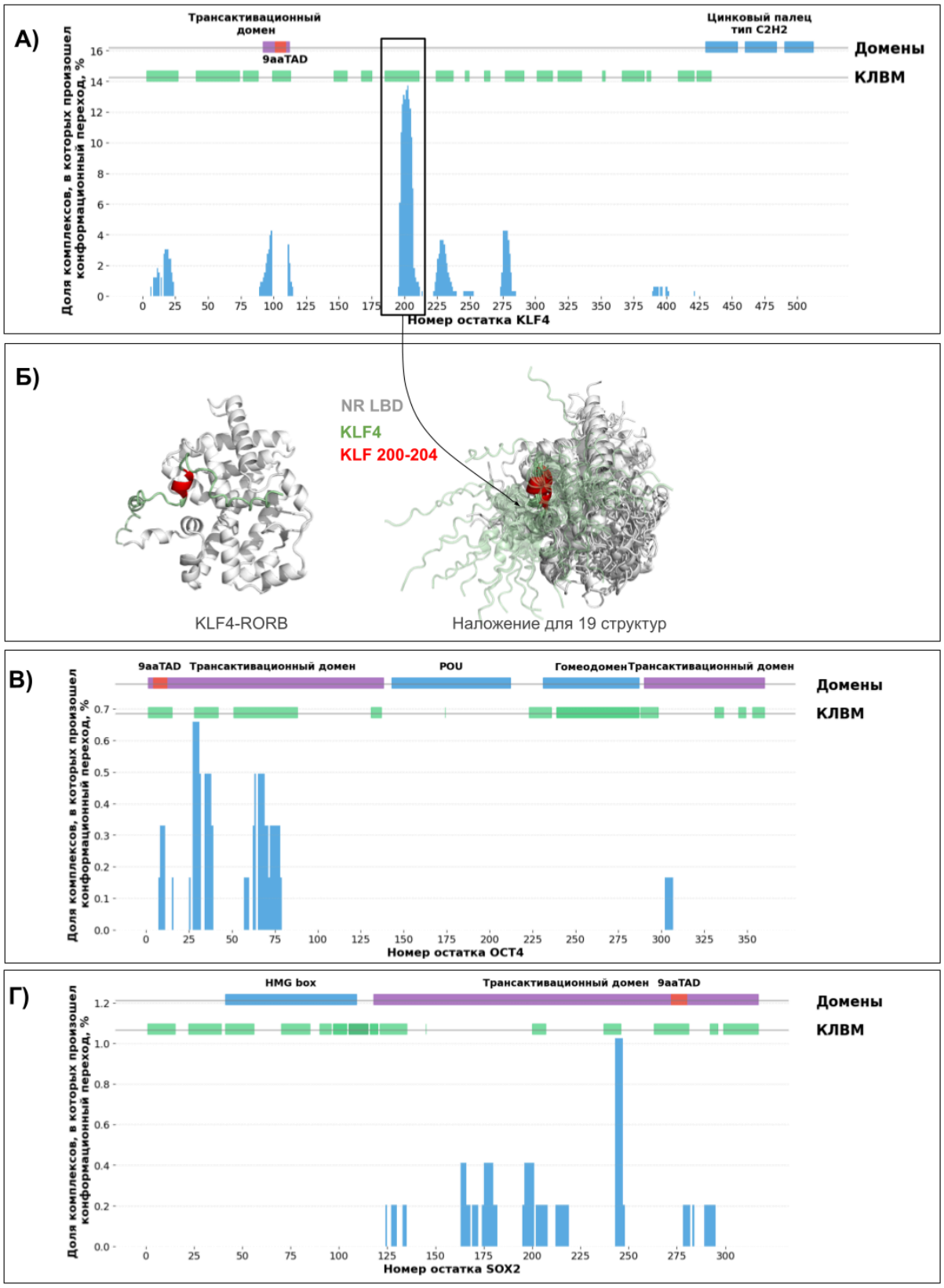
А)
Трансактивационный домен
9aaTAD
Цинковый палец тип C2H2
Домены
КЛВМ
Доля комплексов, в которых произошел конформационный переход, %
Номер остатка KLF4
Б)
NR LBD
KLF4
KLF 200-204
KLF4-RORB
Наложение для 19 структур
В)
9aaTAD
Трансактивационный домен
POU
Гомеодомен
Трансактивационный домен
Домены
КЛВМ
Доля комплексов, в которых произошел конформационный переход, %
Номер остатка OCT4
Г)
HMG box
Трансактивационный домен
9aaTAD
Домены
КЛВМ
Доля комплексов, в которых произошел конформационный переход, %
Номер остатка SOX2

Рисунок 5. А,В,Г) Распределение доли конформационных переходов из неупорядоченного в упорядоченное состояние в KLF4, OCT4 и SOX2, соответственно, в комплексах среднего и высокого качества. Над графиком приведена доменная архитектура. Зеленым цветом окрашены участки, соответствующие КЛВМ из базы данных MobiDB [35]. Б) Предсказанное взаимодействие фрагмента 196-212 KLF4 с лиганд-связывающим доменом ядерных рецепторов (NR LBD). Слева: пример комплекса с белком RORB, справа: наложение всех 19 структур комплексов KLF 200-204 с NR LBD, выровненных по NR LBD. Участки ядерного рецептора вне LBD и KLF4 вне окружения мотива 200-204 не показаны.

Особенно примечателен один из выделенных коротких линейных мотивов, расположенный в регионе аминокислот 196–212 белка KLF4. Из исходного неупорядоченного состояния он переходит в упорядоченное в 27 белково-белковых комплексах. Еще более интересным оказалось другое свойство данного мотива: он проявляет значительную избирательность в выборе партнера взаимодействия, предпочтительно соединяясь исключительно с белками одного определенного класса — ядерными рецепторами гормонов.

Подробный анализ взаимодействий фрагмента ПТФ в позиции 200–204 (последовательность VAELL) с ядерными рецепторами показал, что этот участок связывается с лиганд-связывающим доменом ядерных рецепторов (nuclear receptor ligand binding domain, NR LBD, InterPro: IPR000536), а именно в области активационной функции AF-2 [47] (см. Рисунок 4Б). Область AF-2 представляет собой классический сайт связывания

ко-регуляторов ядерных рецепторов, который аллостерически модулируется связыванием лиганда на противоположной стороне лиганд-связывающего домена [47]. Взаимодействие KLF4 с ядерными рецепторами происходит в конформации, характерной для ко-активаторов: короткий мотив, известный как NR box (консенсусная последовательность LXXLL, в случае KLF4 — VAELL), занимает участок AF-2, формируя так называемую «заряженную защёлку» из петли H12 лиганд-связывающего домена. В отличие от ко-активаторов, ко-репрессоры имеют более длинные мотивы, которые препятствуют образованию этой «заряженной защёлки» [47].

Что касается взаимодействий других исследованных ПТФ с ядерными рецепторами, для SOX2 было обнаружено всего шесть комплексов с ядерными рецепторами, обладающих средним и высоким качеством предсказания. Во всех этих случаях участок AF-2 ядерного рецептора непосредственно связывается с участком SOX2 в позиции 270–278, который входит в состав мотива 9aaTAD. Интересно, что обычно 9aaTAD взаимодействует с ядерными рецепторами не напрямую, а через посредничество ко-активаторов [48]. Примечательно, что этот фрагмент SOX2 обладает аминокислотной последовательностью, характерной для ко-репрессоров — LRDMISMYL, соответствующей консенсусу LXXX(I/L)XXX(I/L) [47], и формирует более протяженную альфа-спираль. В классических ко-репрессорах такая удлиненная спираль препятствует формированию «заряженной защёлки» в AF-2. Однако в наших моделях эти структурные элементы сосуществуют рядом, что, вероятно, отражает неспособность AlphaFold точно смоделировать конформационные перестройки лиганд-связывающего домена NR LBD.

В анализе OCT4 было обнаружено восемь комплексов с ядерными рецепторами, которые характеризуются средним и высоким качеством структурных предсказаний. В четырех из этих комплексов участок AF-2 ядерного рецептора контактирует с фрагментом OCT4 на позициях 32–37 (последовательность PRTLWS), а в трех других — с участком 330–334 (последовательность FTALY). Интересно, что первый из этих регионов помечен в базе MobiDB как потенциальный короткий линейный мотив для взаимодействий. Однако оба этих участка значительно отклоняются от классических консенсусных последовательностей, присущих ко-регуляторам ядерных рецепторов, а само связывание отличается от типичных структурных паттернов ко-регуляторных взаимодействий.

Для точного выяснения молекулярных деталей взаимодействия ПТФ с ядерными рецепторами гормонов, включая роль лигандов, необходимы дальнейшие экспериментальные и вычислительные исследования. Тем не менее, полученные результаты свидетельствуют о вероятности прямого взаимодействия ПТФ с рядом ядерных рецепторов. Особенно выражено это в случае KLF4, связывающегося через нестандартный NR box мотив VXXLL, а также, возможно, в контактах SOX2 через участок 9aaTAD и OCT4 — через уникальные последовательности, которые пока не соотносятся с известными мотивами ко-регуляторов.

4.3. Классификация по функциям белков-партнеров, участвующих в предсказанных белок-белковых взаимодействиях

Мы провели углубленный анализ функциональной специфики белков-партнёров ядра, для которых были получены предсказанные структуры комплексов с пионерными транскрипционными факторами (ПТФ) с средними и высокими показателями качества. В ходе работы выяснилось, что из исходного набора белков ядра и хроматина (3557 белков) около 726 (примерно 20%) демонстрируют способность взаимодействовать как минимум с одним из исследуемых ПТФ. При этом 45 белков (около 10% от набора партнёров) оказались общими для всех трёх изученных ПТФ — SOX2, OCT4 и KLF4 (см. Рисунок 4A).

Далее был проведён анализ обогащения функциональных терминов по категориям «молекулярная функция» и «биологический процесс» на основе данных Генной Онтологии (Gene Ontology) [49]. Для объединённого списка белков-партнёров выявлено 439 значимых терминов, среди которых в топ-20 вошли термины, связанные с распознаванием специфической двухцепочечной ДНК, регуляцией и механизмами транскрипции, ответом на повреждения ДНК, процессингом микроРНК, поддержкой теломерных участков и деацетилированием гистонов. Наиболее выраженное обогащение отмечено в процессах, тесно связанных с функционированием хроматина.

Учитывая эти результаты, мы осуществили более детальный функциональный анализ белков-партнёров с опорой на классификацию белков хроматина SimChrom (см. раздел Методы). Все исследованные белки-партнёры ПТФ были аннотированы согласно SimChrom. Наиболее

представленные группы белков среди партнёров ПТФ включают: транскрипционные факторы (225 белков), ферменты, действующие на ДНК (51), белки, ассоциированные с центромерами (48), белки, участвующие в процессинге РНК (47), белки репарации ДНК (39), компоненты прерибосом (30), ядерные гормональные рецепторы (28) и белки, связанные с теломерами (25) (см. Рисунок 6А).

Отметим, что среди партнёрских белков ПТФ более половины (52%) составляют пионерные транскрипционные факторы согласно классификации SimChrom, почти половина (48%) — белки ядерных гормональных рецепторов, 44% связаны с прерибосомами, а около 40% — с репарацией ДНК. Эти данные подчеркивают широкий спектр функциональных ролей и разнообразие белковых взаимодействий, в которых участвуют ПТФ, отражая их ключевую роль в регуляции генома и поддержании ядерных процессов.

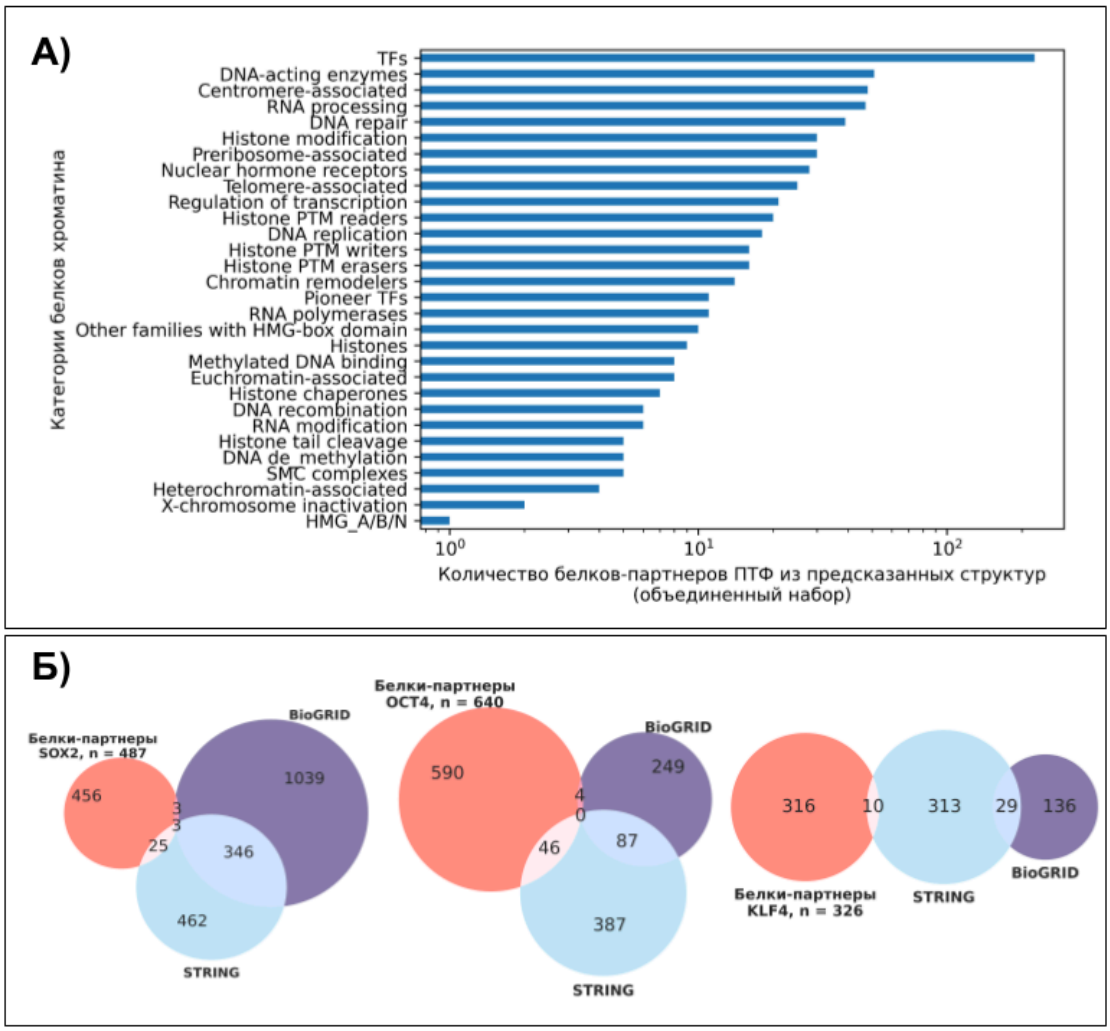

Рисунок 6. А) Принадлежность белков-партнеров ПТФ (объединенный список по всем предсказанным структурам комплексов) категориям белков хроматина по классификации SimChrom. Б) Диаграмма Венна белков-партнеров SOX2, OCT4, KLF4 из предсказанных в данной работе структур, из баз бинарных белок-белковых взаимодействий STRING, BioGRID.

4.4. Сравнение предсказанных комплексов с комплексами из баз данных белок-белковых взаимодействий (STRING, BioGRID) и литературы

Мы сопоставили набор ядерных белков, для которых были получены предсказанные комплексы средней и высокой точности, с партнерскими взаимодействиями, зарегистрированными в базах STRING и BioGRID (см. Рисунок 6Б). Оказалось, что пересечение предсказанных партнёров с данными этих репозиториев составило лишь 6% для SOX2, 7% для OCT4 и 3% для KLF4 от всех моделей со средними и высокими метриками качества. При этом согласованность между самими базами STRING и BioGRID также оказалась ограниченной: общий состав партнёров для каждого ПТФ в обеих базах совпадает только на 32% (SOX2), 12% (OCT4) и 4% (KLF4).

Таким образом, наш предсказательный подход выявил многочисленные новые потенциальные взаимодействия пионерных факторов с ядерными белками, о которых не сообщалось ни в массивах высокопроизводительных экспериментов, ни в литературных обзорах, и лишь незначительная доля этих предсказаний находит подтверждение в существующих базах данных.

Хотя в литературе описаны прямые и косвенные взаимодействия ПТФ с рядом ключевых хроматиновых регуляторов — в частности, с гистон-ацетилтрансферазами CREBBP и EP300, а также с субъединицей SWI/SNF (BRG1) и другими компонентами [50] — подробные структуры этих комплексов до сих пор не установлены. Из обзора 19 таких хроматиновых партнёров [2,4,30,51–55] наша модель AlphaFold 2 Multimer с достаточным качеством (средним или высоким) предсказала лишь семь взаимодействий.

Причём эти комплексы преодолели порог лишь по метрике ipSAE, известной к тому же тенденцией к завышению оценок относительно ipTM и pDockQ (см. Рисунок 2). Такое ограниченное число прямых предсказаний может указывать на то, что взаимодействия ПТФ с указанными хроматиновыми факторами чаще всего опосредованы другими белками или опираются на слабые мультивалентные контакты неупорядоченных участков, что остаётся серьёзным вызовом для алгоритмических методов структурного моделирования.

4.5. Влияние рекуррентной онкологической мутации в KLF4 на белок-белковые взаимодействия

Для изученных пионерных транскрипционных факторов (ПТФ) выявлена лишь одна рецидивирующая точечная онкологическая мутация — замена аминокислоты K409 на Q или N в белке KLF4. Эта мутация зарегистрирована в образцах рака молочной железы, уротелиальной карциномы мочевого пузыря, аденокарциномы предстательной железы (данные cBioPortal [44]), а также в случаях менингиомы согласно литературным источникам [44]. В актуализированной базе данных последовательностей белка KLF4 данный мутационный остаток соответствует позиции 443. Он представляет собой 13-й аминокислотный остаток в первом из трех цинковых пальцев ДНК-связывающего домена и непосредственно контактирует с большой бороздкой ДНК [53]. Кроме того, согласно публикациям, этот участок участвует в связывании с белковым продуктом гена ZNF296 [56].

В предсказанных структурах комплексов обнаружено 23 случая контакта с сайтом мутации K443Q/N (ранее K409Q/N), из которых 4 партнёра относятся к транскрипционным факторам, а 2 — к белкам, ассоциированным с центромерой.

Анализ влияния онкогенной мутации K409Q/N в KLF4 показал среднее значение ΔΔG в +0.379 ккал/моль, что указывает на слабый дестабилизирующий эффект мутаций в среднем. Стандартное отклонение составило 0.621 ккал/моль, отражая значительный разброс влияния мутаций на разные комплексы. Минимальное значение ΔΔG составило -2.312 ккал/моль для комплекса с P55273, где мутация стабилизирует связывание. Максимальное значение ΔΔG составило +1.607 ккал/моль для комплекса с O15397, что свидетельствует о наибольшем дестабилизирующем эффекте. Значение $\Delta\Delta G > 0$ указывает на ухудшение связывания, что наблюдается для большинства комплексов, тогда как $\Delta\Delta G < 0$ говорит о стабилизации связывания, как в случае с P55273. Среднее значение RMSD составило 0.1422 Å, что подтверждает, что мутации не вызывают значительных структурных изменений в белке, и расчеты FoldX проводились на оптимизированных структурах. Среднее значение BSA составило 90.98 Å², что характеризует умеренный контакт между Klf4 и партнерами. Вклад различных энергетических компонентов в ΔΔG приведен в **таблице 3**.

Таблица 3. Вклад различных энергетических компонентов в стабильность мутантных комплексов KLF4

Компонента	**Средний вклад (ккал/моль)**	**Интерпретация**
Силы Ван-дер-Ваальса	+0.6406	Основной **дестабилизирующий** фактор. Указывает на нарушение упаковки.
Водородные связи (осн.)	-0.2210	Слабый **стабилизирующий** эффект за счет взаимодействий с основными цепями.
Водородные связи (бок.)	+0.3951	**Потеря** выгодных взаимодействий с боковыми цепями.
Электростатика	+1.4416	Сильный **дестабилизирующий** вклад (замена заряженного лизина на нейтральные аспарагин/глутамин).

Сольватация (поляр.)	+0.3724	Ухудшение взаимодействий с водой для полярных остатков.
Сольватация (неполяр.)	-0.6070	Компенсаторный **стабилизирующий** эффект для гидрофобных областей.
Энтропия боковых цепей	+0.9232	Увеличение подвижности боковых цепей после мутации (дестабилизация).

Таким образом, мутации K443N/Q в Klf4 в среднем слабо дестабилизируют связывание с белками-партнерами (ΔΔG = +0.38 ккал/моль). Основные причины дестабилизации включают нарушение электростатических взаимодействий (вклад +1.44 ккал/моль) и потерю ван-дер-ваальсовых контактов (+0.64 ккал/моль). Исключение составил комплекс с P55273, который показал стабилизацию (ΔΔG = -2.31 ккал/моль), вероятно, за счет оптимизации гидрофобного ядра. Структурная стабильность поддерживается низким значением RMSD (0.14 Å), что подтверждает, что мутации не искажают глобальную структуру Klf4. Для предсказания функциональных последствий мутаций рекомендуется дополнительный

анализ, например, с использованием молекулярной динамики, а также следует уделить внимание партнерам с наибольшим ΔΔG.

5. Выводы

5.1. Методами генеративного ИИ было предсказано 10 632 комплекса белков ПТФ (SOX2, OCT4, KLF4) человека с другими белками ядра. Из числа предсказанных структур 14% имеют высокое качество, большая часть полученных взаимодействий ранее не была описана в литературе.

5.2. Были выявлены ключевые участки ПТФ в белок-белковых комплексах для которых характерен переход в упорядоченное состояние при связывании. Подтверждены специфические взаимодействия KLF4 с ядерными рецепторами гормонов. Предложен механизм действия онкологической мутации K409Q/N в KLF4: показано, что она слабо дестабилизирует белок-белковые комплексы.

5.3. Выявлено, что среди белков-партнеров ПТФ преимущественно представлены транскрипционные факторы, гистон-модифицирующие ферменты, белки репарации ДНК и ядерные рецепторы.

6. Заключение

Таким образом, в данной работе мы провели масштабный анализ белок-белковых взаимодействий пионерных транскрипционных факторов (ПТФ) человека из «Коктейля Яманаки» — SOX2, OCT4 и KLF4 — с другими ядерными белками с использованием нейросетевых алгоритмов. Использование модели AlphaFold2-Multimer позволило не только построить структурные модели тысяч комплексов, но и провести систематическую оценку достоверности этих предсказаний с помощью совокупности метрик качества. Мы выделили сотни вероятных взаимодействий с высокой степенью достоверности, включая те, которые ранее не были описаны в литературе или обнаружены в экспериментальных базах данных.

Особое внимание было уделено структурному анализу интерфейсов взаимодействия, локализации взаимодействующих доменов, а также аминокислотному составу интерфейсных участков. Выявлена значительная роль ДНК-связывающих доменов в установлении специфичных контактов, а также активное участие трансактивационных доменов и коротких линейных мотивов (в том числе 9aaTAD) в неспецифичных и мультивалентных взаимодействиях. Мы показали, что такие мотивы способны индуцировать переход неупорядоченных участков в упорядоченные состояния, что свидетельствует об их ключевой роли в организации взаимодействий ПТФ с партнерскими белками.

Функциональная классификация белков-партнёров и анализ обогащения по терминам Gene Ontology показали, что большинство выявленных взаимодействий затрагивают ключевые регуляторные процессы ядра: транскрипцию, репарацию ДНК, процессинг РНК и функции

центромеры. Мы идентифицировали белки-партнеры из числа хроматиновых белков, взаимодействующих с ПТФ, о которых ранее не сообщалось в литературе, тем самым расширив текущие представления о молекулярных механизмах действия SOX2, OCT4 и KLF4.

Особое внимание в работе было уделено взаимодействиям KLF4 с ядерными рецепторами гормонов. Был обнаружен неклассический короткий линейный мотив VAELL, демонстрирующий избирательное связывание с лиганд-связывающим доменом ядерных рецепторов, что позволяет предположить новую функцию KLF4 в регуляции гормонозависимой транскрипции. Мы также выявили, что ряд взаимодействий ПТФ локализован в области аминокислотной замены K409Q/N в KLF4 — рекуррентной онкогенной мутации, ассоциированной с несколькими типами рака. Расчёты свободной энергии связывания показали, что эта мутация в ряде случаев ослабляет взаимодействия с белками-партнерами, что потенциально может быть вовлечено в молекулярные механизмы опухолевой трансформации.

Полученные результаты открывают новые перспективы для экспериментального подтверждения предсказанных взаимодействий и для изучения роли ПТФ в регуляции ядерных процессов, включая механизмы репрограммирования, эпигенетической памяти и онкогенеза. Разработанный в рамках данной работы подход может быть масштабирован на другие белковые системы и служить основой для более глубокого понимания регуляторных сетей ядра клетки.

7. Список литературы

1. Iwafuchi-Doi M., Zaret K.S. Pioneer transcription factors in cell reprogramming // Genes Dev. 2014. Vol. 28, № 24. P. 2679–2692.
2. Balsalobre A., Drouin J. Pioneer factors as master regulators of the epigenome and cell fate // Nat. Rev. Mol. Cell Biol. 2022. Vol. 23, № 7. P. 449–464.
3. Li M., Izpisua Belmonte J.C. Deconstructing the pluripotency gene regulatory network // Nat. Cell Biol. 2018. Vol. 20, № 4. P. 382–392.
4. Takahashi K. et al. Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors // Cell. 2007. Vol. 131, № 5. P. 861–872.
5. Wiegand C., Banerjee I. Recent advances in the applications of iPSC technology // Curr. Opin. Biotechnol. 2019. Vol. 60. P. 250–258.
6. Jonas F., Navon Y., Barkai N. Intrinsically disordered regions as facilitators of the transcription factor target search // Nat. Rev. Genet. 2025. Vol. 26, № 6. P. 424–435.
7. Már M., Nitsenko K., Heidarsson P.O. Multifunctional Intrinsically Disordered Regions in Transcription Factors // Chem. – Eur. J. 2023. Vol. 29, № 21. P. e202203369.
8. Udupa A., Kotha S.R., Staller M.V. Commonly asked questions about transcriptional activation domains // Curr. Opin. Struct. Biol. 2024. Vol. 84. P. 102732.
9. Sanborn A.L. et al. Simple biochemical features underlie transcriptional activation domain diversity and dynamic, fuzzy binding to Mediator // eLife. 2021. Vol. 10. P. e68068.
10. Erijman A. et al. A High-Throughput Screen for Transcription Activation Domains Reveals Their Sequence Features and Permits Prediction by Deep

Learning // Mol. Cell. 2020. Vol. 78, № 5. P. 890-902.e6.
11. Soto L.F. et al. Compendium of human transcription factor effector domains // Mol. Cell. 2022. Vol. 82, № 3. P. 514–526.
12. Piskacek M. et al. The 9aaTAD activation domains in the four Yamanaka Oct4, Sox2, Myc, and Klf4 transcription factors essential during the stem cell development: preprint. Genetics, 2019.
13. Orsetti A. et al. Structural dynamics in chromatin unraveling by pioneer transcription factors // Biophys. Rev. 2024. Vol. 16, № 3. P. 365–382.
14. Fedulova A.S. et al. Molecular dynamics simulations of nucleosomes are coming of age // WIREs Comput. Mol. Sci. 2024. Vol. 14, № 4. P. e1728.
15. Luzete-Monteiro E., Zaret K.S. Structures and consequences of pioneer factor binding to nucleosomes // Curr. Opin. Struct. Biol. 2022. Vol. 75. P. 102425.
16. Jumper J. et al. Highly accurate protein structure prediction with AlphaFold // Nature. 2021. Vol. 596, № 7873. P. 583–589.
17. Abramson J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3 // Nature. 2024. Vol. 630, № 8016. P. 493–500.
18. Evans R. et al. Protein complex prediction with AlphaFold-Multimer. 2021.
19. James A.M. et al. In silico screening identifies SHPRH as a novel nucleosome acidic patch interactor. bioRxiv, 2024. P. 2024.06.26.600687.
20. Schmid E.W., Walter J.C. Predictomes, a classifier-curated database of AlphaFold-modeled protein-protein interactions // Mol. Cell. 2025. Vol. 85, № 6. P. 1216-1232.e5.
21. Soufi A., Donahue G., Zaret K.S. Facilitators and Impediments of the Pluripotency Reprogramming Factors' Initial Engagement with the Genome // Cell. 2012. Vol. 151, № 5. P. 994–1004.
22. King H.W., Klose R.J. The pioneer factor OCT4 requires the chromatin

remodeller BRG1 to support gene regulatory element function in mouse embryonic stem cells // eLife. 2017. Vol. 6. P. e22631.

23. Dodonova S.O. et al. Nucleosome-bound SOX2 and SOX11 structures elucidate pioneer factor function // Nature. 2020. Vol. 580, № 7805. P. 669–672.
24. Friman E.T. et al. Dynamic regulation of chromatin accessibility by pluripotency transcription factors across the cell cycle // eLife. 2019. Vol. 8. P. e50087.
25. Maresca M. et al. Pioneer activity distinguishes activating from non‐activating SOX2 binding sites // EMBO J. 2023. Vol. 42, № 20. P. e113150.
26. Chronis C. et al. Cooperative Binding of Transcription Factors Orchestrates Reprogramming // Cell. 2017. Vol. 168, № 3. P. 442-459.e20.
27. Mayran A., Drouin J. Pioneer transcription factors shape the epigenetic landscape // J. Biol. Chem. 2018. Vol. 293, № 36. P. 13795–13804.
28. Zhu Z. et al. Mitotic bookmarking by SWI/SNF subunits // Nature. 2023. Vol. 618, № 7963. P. 180–187.
29. Eberhardt J. et al. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings // J. Chem. Inf. Model. 2021. Vol. 61, № 8. P. 3891–3898.
30. Burke D.F. et al. Towards a structurally resolved human protein interaction network // Nat. Struct. Mol. Biol. 2023. Vol. 30, № 2. P. 216–225.
31. El-Gebali S. et al. The Pfam protein families database in 2019 // Nucleic Acids Res. 2019. Vol. 47, № D1. P. D427–D432.
32. Blum M. et al. InterPro: the protein sequence classification resource in 2025 // Nucleic Acids Res. 2025. Vol. 53, № D1. P. D444–D456.
33. The UniProt Consortium et al. UniProt: the Universal Protein Knowledgebase in 2023 // Nucleic Acids Res. 2023. Vol. 51, № D1. P. D523–D531.

34. Thul P.J. et al. A subcellular map of the human proteome // Science. 2017. Vol. 356, № 6340. P. eaal3321.
35. Li W., Jaroszewski L., Godzik A. Clustering of highly homologous sequences to reduce the size of large protein databases // Bioinformatics. 2001. Vol. 17, № 3. P. 282–283.
36. Zhang J. et al. Computing the Human Interactome. 2024.
37. Steinegger M., Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets // Nat. Biotechnol. 2017. Vol. 35, № 11. P. 1026–1028.
38. Bryant P., Pozzati G., Elofsson A. Author Correction: Improved prediction of protein-protein interactions using AlphaFold2 // Nat. Commun. 2022. Vol. 13, № 1. P. 1694.
39. Dunbrack R.L. Rēs ipSAE loquunt: What's wrong with AlphaFold's ipTM score and how to fix it. bioRxiv, 2025. P. 2025.02.10.637595.
40. Szklarczyk D. et al. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest // Nucleic Acids Res. 2023. Vol. 51, № D1. P. D638–D646.
41. Oughtred R. et al. The BioGRID interaction database: 2019 update // Nucleic Acids Res. 2019. Vol. 47, № Database issue. P. D529–D541.
42. Piovesan D. et al. MOBIDB in 2025: integrating ensemble properties and function annotations for intrinsically disordered proteins // Nucleic Acids Res. 2025. Vol. 53, № D1. P. D495–D503.
43. Buß O., Rudat J., Ochsenreither K. FoldX as Protein Engineering Tool: Better Than Random Based Approaches? // Comput. Struct. Biotechnol. J. 2018. Vol. 16. P. 25–33.
44. Gao J. et al. Integrative Analysis of Complex Cancer Genomics and Clinical

Profiles Using the cBioPortal // Sci. Signal. 2013. Vol. 6, № 269.

45. Fang Z., Liu X., Peltz G. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python // Bioinformatics / ed. Lu Z. 2023. Vol. 39, № 1. P. btac757.
46. Reményi A. et al. Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4 and Sox2 on two enhancers // Genes Dev. 2003. Vol. 17, № 16. P. 2048–2059.
47. Weikum E.R., Liu X., Ortlund E.A. The nuclear receptor superfamily: A structural perspective // Protein Sci. 2018. Vol. 27, № 11. P. 1876–1892.
48. Piskacek M. et al. Nuclear hormone receptors: Ancient 9aaTAD and evolutionally gained NCoA activation pathways // J. Steroid Biochem. Mol. Biol. 2019. Vol. 187. P. 118–123.
49. The Gene Ontology Consortium et al. The Gene Ontology knowledgebase in 2023 // GENETICS / ed. Baryshnikova A. 2023. Vol. 224, № 1. P. iyad031.
50. Barral A., Zaret K.S. Pioneer factors: roles and their regulation in development // Trends Genet. 2024. Vol. 40, № 2. P. 134–148.
51. Jauch R. et al. Crystal Structure and DNA Binding of the Homeodomain of the Stem Cell Transcription Factor Nanog // J. Mol. Biol. 2008. Vol. 376, № 3. P. 758–770.
52. Liu B.H. et al. Targeting cancer addiction for SALL4 by shifting its transcriptome with a pharmacologic peptide // Proc. Natl. Acad. Sci. 2018. Vol. 115, № 30.
53. Megy S. et al. STD and TRNOESY NMR studies for the epitope mapping of the phosphorylation motif of the oncogenic protein β‑catenin recognized by a selective monoclonal antibody // FEBS Lett. 2006. Vol. 580, № 22. P. 5411–5422.

54. Wang L. et al. LIN28 Zinc Knuckle Domain Is Required and Sufficient to Induce let-7 Oligouridylation // Cell Rep. 2017. Vol. 18, № 11. P. 2664–2675.
55. Michael A.K. et al. Mechanisms of OCT4-SOX2 motif readout on nucleosomes // Science. 2020. Vol. 368, № 6498. P. 1460–1465.
56. He Z., He J., Xie K. KLF4 transcription factor in tumorigenesis // Cell Death Discov. 2023. Vol. 9, № 1. P. 118.