«УТВЕРЖДАЮ»:
Заместитель декана биологического факультета
МГУ имени М.В.Ломоносова,
доктор биологических наук,
профессор А.М.Рубцов

«_____» _____ 202__ г.

ЗАКЛЮЧЕНИЕ
кафедры биоинженерии биологического факультета Федерального государственного бюджетного образовательного учреждения высшего образования «Московский государственный университет имени М.В.Ломоносова»

Диссертация «Анализ разнообразия и классификация последовательностей белков гистонов в эукариотах, археях и вирусах» выполнена на кафедре биоинженерии биологического факультета МГУ имени М.В.Ломоносова.

В период подготовки диссертации Сингх-Пальчевская Лавприт обучается в очной аспирантуре биологического факультета на кафедре биоинженерии по специальности 1.5.8 – «Математическая биология, биоинформатика» с 01.10.2019 г., планируемая дата окончания обучения 30.09.2023 г.; а также работает в федеральном государственном бюджетном образовательном учреждении высшего образования «Московский государственный университет имени М.В.Ломоносова» в должности младшего научного сотрудника.

В 2018 г. окончила федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет «Высшая Школа Экономики»» по специальности «Прикладная математика и информатика».

Свидетельство об окончании аспирантуры, подтверждающее сдачу кандидатских экзаменов, будет выдано в 2023 г.

Научный руководитель – д.ф.-м.н., чл.-корр. РАН, профессор РАН Шайтан Алексей Константинович, доцент кафедры биоинженерии биологического факультета федерального государственного бюджетного образовательного учреждения высшего образования «Московский государственный университет имени М.В.Ломоносова».

По итогам обсуждения принято следующее заключение.

1. Диссертационная работа соискателя ученой степени посвящена исследованию разнообразия белков гистонов в широком спектре живых организмов методами биоинформатики. Работа выполнена на современном научно-методическом уровне соответствующем международным стандартам. В работе разработаны новые методики биоинформатического анализа и с их помощью получены новые знания о разнообразии и свойствах белков гистонов.

2. Результаты, изложенные в диссертации, были получены соискателем ученой степени лично.

3. Результаты, полученные соискателем ученой степени, являются достоверными, что подтверждается их апробацией на ведущих российских и международных конференциях, обсуждением на научных семинарах в МГУ, публикацией в рецензируемых материалах и тезисах конференций, в том числе в журналах, входящих в Web of Science.

4. Новизна диссертационной работы и ее результатов заключается в том, что соискателем ученой степени была разработана новая система классификации гистоновых белков, которая учитывает видовое и внутривидовое разнообразие, включая археи и вирусы; охарактеризованы новые потенциально функционально значимые вариации аминокислотных последовательностей гистоновых белков; разработаны новые гибкие алгоритмы автоматической классификации гистоновых белков, позволяющие проводить эволюционный анализ гистонов не только в эукариотах, но и в археях и вирусах; выявлены ранее неизвестные подсемейства коротких гистонов H2A.

5. Результаты проведенных соискателем ученой степени исследований расширяют знания о различных семействах и подсемействах гистоновых белков, что помогает глубже понять функционирование генома как в эукариотах, так и в археях и вирусах. Ввиду того, что многие гистоновые белки играют важную роль в развитии некоторых заболеваний, результаты диссертационной работы могут помочь в решении задач, связанных с изучением механизмов болезней и способов борьбы с ними. Анализ и классификация гистоновых белков вирусов, простейших и паразитов улучшит понимание механизмов вирулентности и разработки методов лечения.

6. Текст диссертации соответствует установленным правилам научного цитирования, библиографические ссылки оформлены корректно.

7. Диссертационное исследование по своему содержанию соответствует заявленной специальности 1.5.8 – «Математическая биология, биоинформатика».

8. Основные идеи и положения работы изложены в 5 научных работах автора общим объемом 0,944 п.л., в том числе 1 публикация (объемом 0,0469 п.л.) в рецензируемых научных изданиях, рекомендованных для защиты в диссертационном совете МГУ по специальности.

9. В своих научных трудах соискатель разработала новые подходы и алгоритмы классификации гистоновых белков, которые позволили выявить новые функционально значимые особенности их подсемейств; написала программы для обновления функционала базы данных HistoneDB, которая позволяет проводить эволюционный анализ последовательностей; выявила ранее неизвестные подсемейства коротких H2A гистонов; а также охарактеризовала их потенциально функционально значимые различия, которые предположительно изменяют физико-химические свойства нуклеосомы.

Диссертация «Анализ разнообразия и классификация последовательностей белков гистонов в эукариотах, археях и вирусах»

Сингх-Пальчевской Лавприт по пунктам 9, 10, 11 и 14 соответствует, а по пункту 13 не соответствует требованиям, установленным в соответствии с Федеральным законом "О науке и государственной научно-технической политике" в Постановлении Правительства РФ "О порядке присуждения ученых степеней" (вместе с "Положением о присуждении ученых степеней"). Представленная диссертация рекомендуется к защите на соискание ученой степени кандидата физико-математических наук по научной специальности 1.5.8 – «Математическая биология, биоинформатика» после доработки (публикации основных научных результатов диссертации в рецензируемых научных изданиях в виде статей в количестве достаточном для представления диссертации к защите в диссертационный совет).

Заключение принято на заседании кафедры биоинженерии биологического факультета МГУ имени М.В.Ломоносова. Присутствовало на заседании 18 чел. Результаты голосования: «за» - 18 чел., «против» - 0 чел., «воздержалось» - 0 чел., протокол № 9 от «17» августа 2023 г.

Зам. заведующего кафедрой биоинженерии
Биологического факультета
МГУ имени М.В.Ломоносова
д.ф.-м.н., профессор К.В. Шайтан

МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М.В. ЛОМОНОСОВА
БИОЛОГИЧЕСКИЙ ФАКУЛЬТЕТ

На правах рукописи

Сингх-Пальчевская Лавприт

Анализ разнообразия и классификация последовательностей белков гистонов в эукариотах, археях и вирусах

Специальность 1.5.8 —
«Математическая биология, биоинформатика»

ДИССЕРТАЦИЯ

на соискание ученой степени

кандидата физико-математических наук

Научный руководитель:
доцент кафедры биоинженерии
доктор физико-математических наук
Шайтан А.К.

Москва – 2023

Оглавление

Введение

Актуальность темы исследования. Исследование процессов жизнедеятельности клетки продолжается уже более 50 лет. Однако, несмотря на наличие немалого запаса знаний, мы все еще далеки от комплексного понимания функционирования генома. Поэтому данная проблема является актуальной и в 21 веке.

Регуляция активности генов происходит в ядре клетки и зависит не только от последовательности ДНК. Существует множество разнообразных эпигенетических механизмов, которые оказывают влияние на активность определенных участков генома, не изменяя при этом первичную структуру ДНК. Центральное место в эпигенетических процессах занимают белки гистонов. За счет плотной ассоциации с ДНК и различных пост-трансляционных модификаций аминокислот они влияют на компактизацию и динамику хроматина в ядре клетки.

Важной особенностью белков гистонов является их функциональное и видовое разнообразие. Все эукариотические гистоны можно разделить на пять основных типов: H1 (у птиц называемый H5), H2A, H2B, H3, H4. Известно также, что различные гены паралоги, кодирующие белки одного типа, отличаются физико-химическими свойствами, что приводит к формированию продуктов, отличающихся как структурно, так и функционально. Интересно отметить, некоторые гистоны участвуют в тонкой регуляции работы целых органов. Например, нокаут генов некоторых вариантов гистонов H2B в мышах приводит к ухудшению памяти и обоняния. Гистоновые белки могут быть специфичны для определенных видов организмов или типов ткани. Продолжительное время считалось, что белки гистонов присутствуют исключительно в эукариотической клетке. Однако, около 10 лет назад гистоны были найдены в археях, а относительно недавно обнаружены вирусные гистоны. На данный момент нам известно очень мало об их структурных и функциональных особенностях. При этом исследования

демонстрируют наличие значимых структурных отличий в сравнении с эукариотическими гистонами.

Отсутствие систематизированных знаний о различных семействах гистоновых белков и их особенностях приводит к тому, что задача классификации гистонов является нетривиальной. Чтобы провести комплексный анализ разнообразия белков гистонов в различных живых организмах, необходимо иметь набор всех известных аминокислотных последовательностей гистонов. Однако, различные белковые базы данных содержат не только гистоновые, но и огромное количество последовательностей малоизвестной природы.

В 2015 году была разработана база данных HistoneDB 2.0, которая включала около 80 тысяч гистоновых белков. Алгоритмы автоматической классификации, лежащие в ее основе, позволяют отличить аминокислотную последовательность гистонов от других белков, а также определить подсемейство, к которому принадлежит гистоновый белок. База данных HistoneDB оснащена множеством инструментов для сравнения и анализа аминокислотных последовательностей гистонов. Она широко используется исследователями разных стран и является одним из важных инструментов для анализа гистоновых белков. Однако, за последние 8 лет появилось множество новых знаний. Различные литературные источники свидетельствуют о нахождении новых функционально значимых подсемейств гистонов у разных видов живых организмов, включая археи и вирусы. Известно, что их аминокислотная последовательность может отличаться, как значительно, так и всего несколькими аминокислотами. Вдобавок, они могут характеризоваться специфическими мотивами или дополнительными доменами. Вследствии этого автоматическое обнаружение новых подсемейств и их особенностей требует модернизации алгоритмов. Более того, произошел ряд технических изменений, которые могут повлиять стабильность работы функциональных инструментов базы данных.

Таким образом, изучение разнообразия гистоновых белков и модернизация алгоритмов поиска и классификации гистонов являются актуальными задачами 21 века.

Степень разработанности темы. Впервые гистоновые белки были обнаружены в 1884 году. Немецкий биохимик и физиолог Альбрехт Коссель продемонстрировал, что “нуклеин”, открытый Фридрихом Мишером в 1869 году, состоит из белкового и небелкового компонента, последний из которых он выделил и описал как совокупность соединений: аденин, цитозин, гуанин, тимин и урацил. Несмотря на такое открытие, основное внимание ученых долго время было приковано к соединению, известному сегодня, как молекула ДНК. Интерес к гистоновым белкам появился в 1960-х годах. Ряд исследований позволил установить химический состав белков хроматина и провести первую классификацию гистоновых белков. Первоначально имена были присвоены следующие: F3, F2A1, F2A2, F2B и F3. В то же время были проведены исследования по анализу взаимодействия гистоновых белков с ДНК, а также была определена роль гистонов в структуре хроматина.

За последние 60 лет обнаружено более 50 различных подсемейств гистоновых белков. Кроме того, за прошедшее десятилетие гистоновые белки были замечены у таких организмов, как археи и вирусы. Подробный обзор литературы, содержащих исследования гистоновых белков и их структурных и функциональных отличий, представлен в главе 1 “Гистоны и их разнообразие”.

В 1995 году была разработана первая база данных Histone Sequence Database, которая предоставляла актуальную информацию обо всех белках хроматина. По мере появления новых знаний она претерпевала различные изменения. Сегодня данная база данных носит название HistoneDB 2.0 и содержит более 80 тысяч аминокислотных последовательностей гистоновых белков. Обзор по разработке и модернизации баз данных гистоновых белков представлен в главе 2 “Базы данных гистоновых белков”.

Цель и задачи исследований. Основными целями исследования являются изучение, анализ и систематизация разнообразия и классификация аминокислотных последовательностей гистоновых белков в эукариотах, археях и вирусах.

Для достижения данной цели были поставлены следующие задачи:

1. формирование нового курируемого набора аминокислотных последовательностей гистоновых белков на основе комплексного анализа литературы и биоинформатических баз данных;
2. анализ разнообразия и вариаций аминокислотных последовательностей белков гистонов курируемого набора;
3. расширение и модернизация базы данных HistoneDB 2.0 для получения более полного набора аминокислотных последовательностей гистоновых белков:
   - 3.1. техническое обновление, оптимизирующее скорость и обеспечивающее стабильность функциональных инструментов базы данных;
   - 3.2. обнаружение несовершенств автоматической классификации с помощью статистического анализа разнообразия всех последовательностей гистоновых белков, хранящихся в базе данных HistoneDB 2.0, для разработки улучшенных алгоритмов автоматической классификации гистонов;
   - 3.3. автоматизация иерархического процесса построения множественных выравниваний для получения выравниваний аминокислотных последовательностей различных функционально значимых семейств и подсемейств гистонов широким эволюционным разнообразием;
   - 3.4. разработка усовершенствованного алгоритма автоматического поиска и классификации аминокислотных последовательностей гистоновых белков;

3.5. разработка нового интерактивного веб-интерфейса, который позволяет взаимодействовать с обновленной базой данных для обнаружения новых функционально значимых подсемейств гистонов и их особенностей, включая специфичные сайты, влияющие на структуру нуклеосомы;

4. поиск новых новых функционально значимых подсемейств гистонов и их особенностей, включая специфичные сайты, влияющие на структуру нуклеосомы.

Объем и структура работы. Диссертация состоит из введения, четырех глав и заключения. В первой главе представлен обзор литературы, посвященной исследованиям гистоновых белков, их свойств и особенностей, обсуждается разнообразие гистонов в различных видах живых организмов и их роль в функционировании генома, а также представлен обзор существующих баз данных, в которых собраны аминокислотные последовательности гистоновых белков. Во второй главе описаны методы, применяемые для анализа разнообразия и классификации гистоновых белков. Третья глава посвящена разработке новых более точных и гибких алгоритмов поиска и классификации гистоновых белков, а также расширению и модернизации базы данных HistoneDB 2.0. И наконец, в четвертой главе представлены результаты анализа разнообразия гистоновых белков в широком спектре живых организмов, а также их влияние на структурные свойства нуклеосомы. Список литературы содержит 51 работу. Полный объем научной работы составляет 97 страниц, включая 14 рисунков, 2 таблицы и приложения.

Научная новизна работы. Гистоновые белки обладают очень широким видовым и функциональным разнообразием. Различные особенности их первичных и вторичных структур способны влиять на жизненно важные функции организма. Несмотря на накопленные знания о гистоновых белках, их семействах и подсемействах, мы пока далеки от целостного понимания их многообразия в широком спектре живых организмов.

В данной исследовательской работе проведен комплексный анализ гистоновых белков в эукариотах, археях и вирусах. Для этого сформирован новый курируемый набор аминокислотных последовательностей гистоновых белков, а также разработана новая иерархическая система классификации их на различные функционально значимые семейства и подсемейства. В результате биоинформатического анализа аминокислотных последовательностей гистонов в широком спектре живых организмов выявлены ранее неизвестные подсемейства, а также охарактеризованы их особенности, физико-химические и структурные свойства, влияющие на стабильность нуклеосомы. Кроме того, выполнена модернизация широко применяемой базы данных HistoneDB 2.0. Разработаны новые более точные и гибкие алгоритмы автоматической классификации, которые позволили расширить разнообразие аминокислотных последовательностей гистоновых белков, хранящихся в ней, а также обнаружить новые функционально значимые подсемейства. Обновленная база данных HistoneDB 3.0 содержит более 186000 аминокислотных последовательностей гистоновых белков и более 50 различных гистоновых подсемейств.

Теоретическая и практическая значимость работы. Результаты данной научной работы могут применяться в фундаментальных и прикладных исследованиях в области молекулярной биологии, связанные с изучением функционирования генома. Понимание разнообразия и классификация гистоновых белков в эукариотах, археях и вирусах позволит приблизиться к более комплексному знанию о различных эпигенетических механизмах, регулирующие активность генов, глубже оценить значимость и роль гистонов в таких процессах, как транскрипция, репликация и репарация, а также построить более строгие модели регуляции работы генома в различных видах живых организмов.

Ввиду того, что многие гистоновые белки играют важную роль в развитии некоторых заболеваний, систематизация знаний о функционально значимых подсемействах гистоновых белков будет значима и для прикладных

задач, связанных с изучением механизмов болезней и способов борьбы с ними. Кроме того, анализ и классификация гистоновых белков вирусов, простейших и паразитов улучшит понимание механизмов вирулентности и разработки методов лечения.

Методология и методы исследования. В данной научной работе для проведения комплексного анализа разнообразия аминокислотных последовательностей гистоновых белков в широком спектре живых организмов применяются методы биоинформатики, вычислительной биологии и молекулярного моделирования. Для реализации задач, поставленных в рамках исследования, были написаны скрипты с использованием прикладных библиотек высокоуровневого языка программирования Python 3, в число которых вошли biopython, pytexshade, pynucl, matplotlib, seaborn, scipy, numpy, pandas, ete3, Django и др.

Важным методологическим инструментом данной научной работы являются методы построения множественных выравниваний CLUSTALW2 и MUSCLE, которые применяются как для изучения разнообразия и вариаций аминокислотных последовательностей белков гистонов нового курируемого набора, так и для разработки новых алгоритмов автоматической классификации гистонов. Программное обеспечение MUSCLE также использовалось при разработке автоматизированного подхода для иерархического построения множественных выравниваний, в результате которого были получены выравнивания аминокислотных последовательностей различных функционально значимых семейств и подсемейств гистонов в эукариотах, археях и вирусах.

На основе множественных выравниваний был проведен филогенетический анализ с использованием программного обеспечения PhyML, в основе которого лежат алгоритмы Neighbor-Joining и поиск максимального правдоподобия. Для обнаружения новых функционально значимых подсемейств гистоновых белков, была проведена кластеризация на основе метода UPGMA.

В ходе статистического анализа разнообразия всех последовательностей гистоновых белков, хранящихся в базе данных HistoneDB 2.0, для обнаружения несовершенств были построены множественные и глобальные выравнивания с использованием MUSCLE и алгоритмов динамического программирования Нидлмана-Вунша. Для расширения базы данных HistoneDB и улучшения алгоритмов автоматического поиска и классификации гистоновых белков применялись программы для поиска гомологичных белков, такие как HMMER и BLASTP.

Для обеспечения стабильной работы функциональных элементов базы данных HistoneDB 3.0, а также ускорения процесса ее автоматического наполнения, использовались Django Framework и MySQL, реализующие веб-приложение и реляционную систему управления базами данных, соответственно. С целью улучшения устойчивости и переносимости программной системы были собраны Docker-образы и Docker-контейнеры для запуска базы данных.

Положения, выносимые на защиту.

1. Разработанная новая система классификации гистоновых белков, учитывающая, как видовое, так и внутривидовое разнообразие, позволяет выявлять особенности последовательности гистонов, потенциально функционально значимые для различных подсемейств гистонов.
2. Разработанная база данных HistoneDB 3.0 позволила выявить и охарактеризовать гистоновые и гистоноподобные белки всех известных на данных момент белковых последовательностей, в результате была создана база данных белков гистонов размером более 186 тысяч последовательностей, которая позволяет проводить эволюционный анализ последовательностей гистонов в широком спектре живых организмов.

3. Продемонстрировано, что несмотря на то, что многие гистоновые белки являются эволюционно высоко консервативными, среди них встречаются белки, у которых идентичность аминокислотных последовательностей в области глобулярных доменов ниже 46%, а вариации в специфичных сайтах предположительно изменяют физико-химические свойства нуклеосомы.
4. Охарактеризовано подсемейство коротких вариантов H2A, в котором выявлены ранее неизвестные подсемейства, обладающие идентичностью аминокислотных последовательностей не менее 52%, а также отличающиеся остатками, расположенными в области кислотного лоскута.
5. База данных HistoneDB 3.0 является многофункциональным инструментом, который позволяет сравнивать, анализировать и классифицировать аминокислотные последовательности гистонов, а также комплексно оценить разнообразие гистоновых белков в эукариотах, археях и вирусах.
6. Благодаря модернизации алгоритмов базы данных HistoneDB 3.0 модели отбора и классификации стали более точными и гибкими для определения различий между подсемействами гистоновых белков в широком спектре живых организмов.

Степень достоверности и апробация результатов. Достоверность полученных результатов обеспечивается их апробацией на ведущих российских и международных конференциях и публикацией в рецензируемых журналах и сборниках конференций, в том числе в журналах, входящих в Web of Science. Результаты находятся в соответствии с результатами, полученными другими авторами. Материалы диссертационной работы докладывались и обсуждались на научных семинарах в МГУ.

Личный вклад. Основная методология и результаты исследований, изложенные в диссертации, получены автором лично. Вклад автора заключается в формировании нового курируемого набора аминокислотных

последовательностей гистоновых белков, в проведении биоинформатического анализа разнообразия белков гистонов в широком спектре живых организмов, в разработке новых автоматических алгоритмов поиска и классификации аминокислотных последовательностей гистоновых белков, а также в расширении и модернизации базы данных HistoneDB 2.0.

Благодарности. Работы, описанные в данной диссертации, поддерживаются грантами РНФ №18-74-10006 и президента РФ МД-1131.2022.1.4.

Автор выражает благодарность своему научному руководителю А.К. Шайтану, а также своим коллегам, которые помогали с решением некоторых задач для расширения и модернизации базы данных HistoneDB, А.К. Грибковой и А.М. Неугодову.

Глава 1. Белки гистоны и их разнообразие

1.1 Введение

Ключевую роль в развитии, жизнедеятельности и адаптации живых организмов к окружающей среде играет функционирование их генома. Сегодня известно, что молекула ДНК содержит полноценную инструкцию по всем процессам, необходимым для роста и деления клетки. Формирование клетки начинается с репликации, в ходе которой образуются две дочерние ДНК на основе родительской. Каждый признак или функция организма закодированы в виде структурной единицы генома, именуемой геном. Конечным продуктом гена является белок, выполняющий определенные функции. В то же время известно, что не весь геном способен кодировать белки. Часть его является некодирующей и существует лишь для того, чтобы регулировать транскрипцию генов. Более того, кодирующие гены могут быть активными или нет в разные моменты времени или в зависимости от типа клетки.

Тонкая взаимная регуляция активности генов, осуществляемая в ядрах клеток живых организмов — ключевой процесс, позволяющий организмам функционировать и развиваться. Существует ряд процессов и факторов, влияющих на изменение активности генов, среди которых ключевую роль занимают эпигенетические механизмы. Физико-химические свойства молекулы ДНК, а также ее взаимодействия с различными белками, приводят к формированию различных модификаций ДНК, ремоделированию хроматина и другим клеточным процессам, которые меняют структуру и динамику хроматина. Эпигенетические механизмы регулируют синтез белков, не изменяя первичную структуру ДНК. Благодаря изменениям в структуре хроматина, некоторые гены становятся доступными, или наоборот недоступными, для различных транскрипционных факторов.

Центральное место среди эпигенетических факторов занимают белки гистоны. Упаковка генома в ядре клетки происходит путем наматывания молекулы ДНК на комплекс, состоящий из белков гистонов. Ввиду того, что

эти белки обычно плотно ассоциированы с ДНК, они способны осуществлять компактизацию ДНК и одновременно регулировать работу генов. Гистоны могут подвергаться посттрансляционным модификациям и ассоциироваться с различными транскрипционными факторами, тем самым влияя на динамику хроматина и регулируя доступность определенных участков генома.

Подавляющее большинство гистонов, именуемые каноническими, участвуют в упаковке ДНК, синтезированной в результате репликации, и экспрессируется преимущественно в ходе S-фазы клеточного цикла [1]. Другие гистоновые белки представляют собой различные варианты гистонов, которые заменяют канонические на протяжении всего клеточного цикла, регулируя тем самым работу генов [2]. В табл. 1 приведены ключевые особенности канонических и вариантных гистонов в эукариотах.

Clustered (canonical) histones	**Histone variants**
экспрессируются преимущественно в S-фазе клеточного цикла	экспрессируются на протяжении клеточного цикла для выполнения определенных функций
участвуют в формировании структуры хроматина во время деления эукариотической клетки	отвечают за динамику хроматина эукариотической клетки (придают нуклеосомам различные структурные свойства, оборачивая большее или меньшее количество ДНК или изменяя стабильность нуклеосом)
гены присутствуют в геноме в нескольких копиях и формируют тандемно повторяющиеся кластеры (самый большой кластер этих генов, называемый HIST1 и	гены не образуют кластеров, разбросаны по всему геному

состоящий из 55 генов, обнаружен у человека и находится на хромосоме 6)	
мРНК большинства белков не полиаденилирована, а вместо поли(А)-хвоста присутствует шпилька	мРНК полиаденилирована
в генах отсутствуют интроны	в генах нередко присутствуют интроны (транскрибируемая с них РНК полиаденируется)

Табл. 1. Отличия канонических и вариантных гистонов в эукариотах

Все гистоновые белки можно разделить на пять основных типов (семейств): коровые H2A, H2B, H3, H4, и линкерный H1 (у птиц называемый H5). По две копии каждого корового гистона (два димера H2A-H2B и тетрамер H3-H4) и ДНК длиной примерно 147 п.о., которая оборачивается вокруг комплекса гистонов, объединяются в структуру, именуемую нуклеосомой (рис. 1) [3]. Каждый гистон состоит из двух структурно важных компонентов: глобулярный домен (histone fold domain, HFD) и хвосты (рис. 2). Глобулярный домен представляет собой структурный мотив, который состоит из трех α-спиралей, соединенных двумя петлями. Исключением являются только пекарские дрожжи, которые имеют два последовательно идущих глобулярных домена, разделенных положительно заряженным карбоксиконцевым доменом (рис. 2). HFD играет важную роль в гистон-гистоновых и гистон-ДНК взаимодействиях и поддерживает структуру нуклеосомы [3]. Хвосты гистоновых белков, выступающие за пределы нуклеосомы, богаты основ́ными аминокислотами, имеют динамичную структуру и подвержены многочисленным посттрансляционным модификациям (ПТМ), включая ацетилирование, метилирование и

фосфорилирование [4]. Их последовательность очень вариативна даже внутри одного подсемейства. Хвосты гистонов влияют на динамику хроматина и регулируют различные эпигенетические механизмы, в том числе транскрипцию, репликацию, рекомбинацию и репарацию ДНК [4].

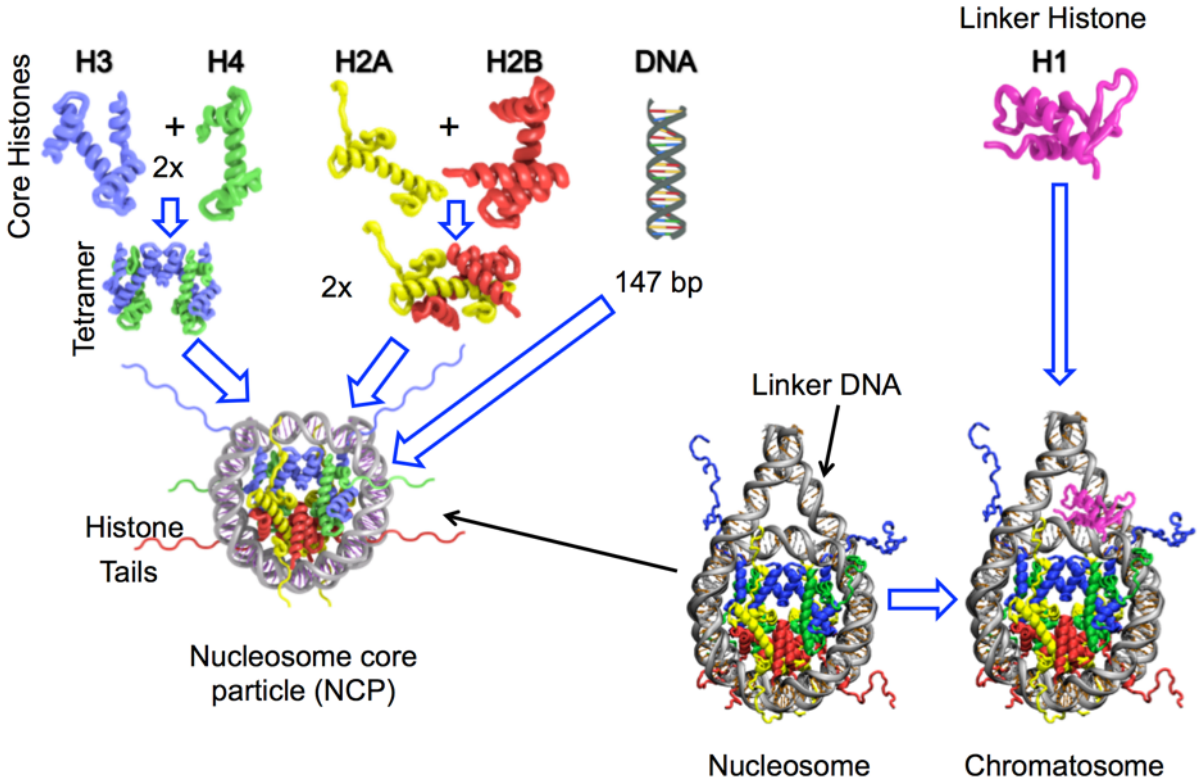

Рисунок 1. Структура нуклеосомы. Восемь гистонов (два димера H2A-H2B и тетрамер H3-H4) формируют коровую часть нуклеосомы (nucleosome core particle, NCP), вокруг которой оборачивается примерно 145-147 п.о. ДНК. Хвосты гистонов обычно выступают за пределы нуклеосомного комплекса. Линкерный гистон H1 фиксирует компактизацию нуклеосомы за счет взаимодействия с линкерной ДНК.

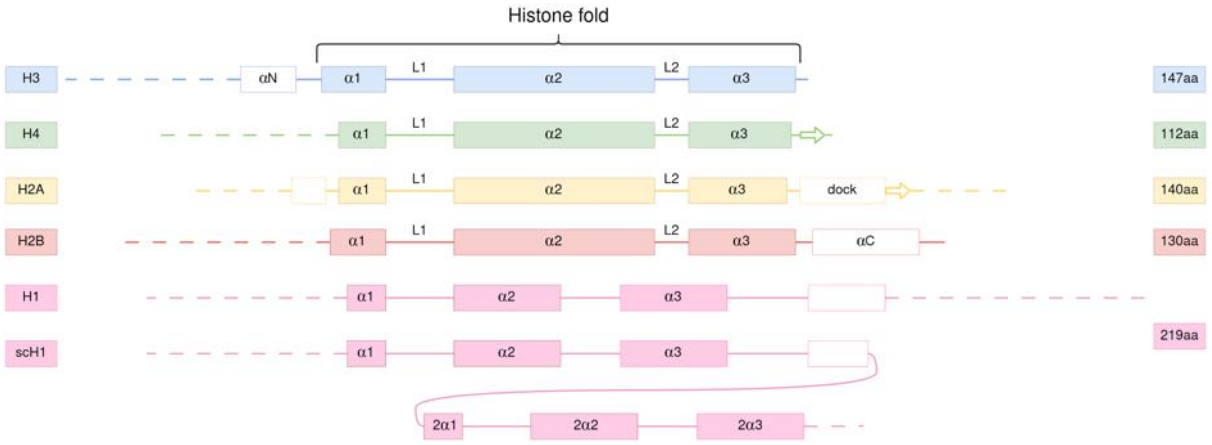

Рисунок 2. Схематическое представление структуры гистонового белка для каждого из пяти семейств. Справа приведены значения средней длины аминокислотной последовательности семейства.

Каждый из пяти типов гистонов, кроме H4, имеют различные варианты (подсемейства), которые замещают канонические гистоны для выполнения специальных функций. Отличия между вариантами и их каноническими формами могут быть как незначительными, затрагивающими лишь несколько аминокислот с сохранением большей части структурных признаков, так и настолько существенными, что они могут быть сравнимы с различиями между типами гистонов (менее 25% идентичности) [2]. Например, гистоновый вариант H3.3 отличается от канонического H3 только несколькими аминокислотными остатками и имеет преимущественно схожие структурные особенности [5]. С другой стороны, белок CENP-A, являющийся специфическим для центромер гистоновым вариантом cenH3, имеет расширенную L1-петлю, и его N-концевой хвост сильно отличается от других вариантов H3 [6].

Разнообразие гистоновых белков отражается как на первичной, так и на вторичной структурах. С одной стороны, гистоны могут отличаться общей длиной и идентичностью по всей последовательности, могут иметь специфические вариации, вставки или делеции аминокислот в определенных позициях, которые в первую очередь затрагивают хвосты. С другой стороны, перечисленные изменения влияют на структуру, физико-химические и

функциональные свойства белка, что в свою очередь влияет на межмолекулярные взаимодействия в хроматине, а также взаимодействия гистонов с транскрипционными факторами. Например, гистоновый белок растений H2A.W имеет специфичный мотив SPKK на С-концевом хвосте, благодаря чему обладает повышенной активностью связывания с малой бороздкой, преобладает в гетерохроматине, участвует в снижении активности генов и отвечает за повреждения ДНК [7].

Некоторые гистоновые варианты специфичны к отдельным таксономическим группам или клеточным культурам. Это можно увидеть на таких примерах, как H2A.W, H2A.M, gH2A, gH2B, которые обнаружены исключительно в растениях, семейство коротких (short) H2A, экспрессирующихся в семенниках, или scH1, характерный пекарским дрожжам.

Отдельного внимания заслуживают гистоны архей и вирусов, которые были обнаружены сравнительно недавно. Их структура и функциональная значимость сильно отличается от эукариотической. Например, гистоны архей способны формировать комплексы “бесконечной” длины, именуемые гипернуклеосомой [8]. А геном вирусов способен кодировать гистоны, состоящие из двух, трех и более субъединиц, которые схожи по аминокислотной последовательности и по вторичной структуре с эукариотическими гистоновыми белками [9]. Несмотря на то, что гистоны и архей, и вирусов эволюционно связаны с эукариотическими, они также разнообразны нуждаются в собственной системе классификации.

В течение долгого времени предполагалось, что гистоновые белки медленно эволюционируют, а большая часть различий между ними приходится на хвосты, разнообразие модификаций которых обуславливает различные регуляторные процессы генома. Сегодня же известно, что гены паралоги, кодирующие белки одного семейства, могут различаться по физико-химическим свойствам, которые используются для регуляции работы генома путем таргетного связывания гистоновых вариантов с определенными

участками генома. Более того, существуют примеры, которые демонстрируют участие некоторых генов в регуляции работы целых органов. Например, нокаут генов некоторых вариантов гистонов H2B в мышах приводит к ухудшению памяти и обоняния.

1.2 Разнообразие гистоновых белков

При каждом делении клетки происходит не только репликация ДНК, но и синтез белков, которые называют каноническими гистонами, чтобы обеспечить упаковку новой ДНК. В клетках млекопитающих для этого требуется синтез около 108 молекул каждого из четырех типов гистоновых белков [10]. Гены канонических гистонов кодируют мРНК, которая отличается от других эукариотических мРНК. Известно, что у эукариот это единственная мРНК, не являющаяся полиаденилированной [11]. На месте ее поли(А)-хвоста обнаружена шпилька, которая играет важную роль в регуляции гистонов. Экспрессия генов канонических гистонов продолжается преимущественно в ходе S-фазы клеточного цикла: вместе с завершением репликации разрушается и мРНК [12]. Поэтому их зачастую называют репликативно-зависимыми генами.

Известно, что гены канонических гистонов не содержат интронов. Поэтому расщепление формирующегося транскрипта с формированием 3'-конца мРНК канонических гистонов является единственным событием процессинга.

Несмотря на то, что экспрессия канонических гистонов часто строго регулируется, механизмы могут различаться. Например, некоторые из генов могут продуцировать полиаденилированные мРНК. В результате анализа глобальной экспрессии генов в нормальных неделящихся тканях было обнаружено подмножество из десяти репликативно-зависимых генов гистонов человека, которые продуцируют полиаденилированные мРНК во всех исследованных неделящихся тканях [10,13]. Также выявлены отличия на уровне видов живых организмов: мРНК канонических гистонов H2A растений и большинства одноклеточных эукариот полиаденилированы [14].

Гены канонических гистонов обычно формируют кластеры, которые содержат несколько копий генов, кодирующих все пять различных типов гистонов. В связи с этим их часто называют кластерными гистонами. При этом организация их генов различна. В двух исследованиях было продемонстрировано, что можно выделить два типа кластеров: 1) тандемно повторяющиеся наборы генов, которое соответствует стандартному представлению, где каждая повторяющаяся единица содержит по одной копии гена гистона, и 2) “рассеянные” кластеры, в которых гены не подчиняются строгому порядку, а перемешаны внутри кластера [12,15]. Единственным исключением является ген человеческого гистонового белка семейства H4. Так же как и все канонические гистоны, он экспрессирует неполиаденилированную мРНК в S фазе. Не смотря на это, он расположен вне кластеров, в которые сгруппированы гены канонических гистонов [12].

На данный момент нам известно четыре дискретных локуса у млекопитающих. Эти кластеры являются “рассеянными” [12]. В геноме человека самый большой кластер (HIST1) находится на хромосоме 6 и содержит более 60 генов, а второй кластер (HIST2) на хромосоме 1 содержит 10–12 генов. Есть 4 гена в третьем отдельном локусе (HIST3) на хромосоме 1 и один репликативно-зависимый ген гистона H4 (HIST4) на хромосоме 12 (с соседним геном H2A, для которого зависимость от репликации неясна). Эта геномная организация консервативна, и все четыре локуса являются синтеническими у млекопитающих [10].

У других позвоночных, в том числе у многих рыб (например, рыбки данио) и земноводных (например, Xenopus), наблюдаются тандемно-повторяющиеся кластеры [10,12]. Например, у курицы присутствует всего один большой кластер, который содержит гены всех пяти типов гистонов [10]. В то же время, известно, что организмы представленных видов хранят в яйце большое количество мРНК гистонов и белков и начинают свое развитие с серии быстрых клеточных циклов в отсутствие зиготическая транскрипция. Предполагается, что такая организация и

количество копий генов гистонов необходимо для обеспечения синтеза большого количества мРНК и белка за короткий период времени [10].

На сегодняшний день известно 18 генов H2A у человека, которые кодируют канонические гистоны H2A. Они представлены в трех кластерах. В самом большом кластере (HIST1), расположенном на хромосоме 6, находятся 13 кодирующих генов (H2AC1, H2AC4, H2AC6, H2AC7, H2AC8, H2AC11, H2AC12, H2AC13, H2AC14, H2AC15, H2AC16, H2AC17) и несколько псевдогенов. В одном из кластеров (HIST2), расположенных на хромосоме 1, находятся 4 кодирующих гена (H2AC18, H2AC19, H2AC20, H2AC21), а другом кластере (HIST3) - 1 кодирующий ген (H2AC25). Аналогично, репликативно-зависимые гены H2B обнаружены в трех локусах человеческого генома: кластер HIST1 кодирует 15 генов H2B (H2BC1, H2BC3-H2BC15, H2BC17) и несколько псевдогенов, HIST2 кодирует 2 гена (H2BC18, H2BC21) и два псевдогена (H2BC19P, H2BC20P), HIST3 кодирует один ген (H2BC26) и один псевдоген (H2BC27P). Один дополнительный ген H2BC12L представлен специфичной для человека дупликацией гена H2BC12 с хромосомы 6 на хромосому 21. Всего для человека известно 19 генов H2B, которые кодируют канонические гистоны H2B. Гены семейства H3 обнаружены в двух кластерах человеческого генома. Из них 10 генов, кодирующих 1 изоформу, расположены на хромосоме 6: H3C1-H3C4, H3C6-H3C8, H3C10-H3C12. 3 гена, кодирующих вторую изоформу, расположены на хромосоме 1: H3C13, H3C14, H3C15 [16].

Как и у человека, репликативно-зависимые гены мыши расположены в трех кластерах. Самый большой из них состоит более, чем из 50 генов, и располагается на хромосоме 13. Два других кластера, поменьше, находятся на хромосомах 3 и 11 [16]. В настоящее время в геноме мыши аннотировано 18 репликативно-зависимых генов H2A: 13 генов в кластере на хромосоме 13, 4 гена в кластере на хромосоме 3 и 1 ген на хромосоме 11. Однако из всех изоформ, описаны только две. Они кодируются генами H2bc1 (ранее

Hist1h2ba, Th2b), расположенном на хромосоме 13, и H2bc21, расположенном на хромосоме 3 [16].

Канонические гистоны H2A, H2B, H3 и H4 образуют фундаментальную единицу хроматина - нуклеосому. Благодаря рентгеновской кристаллической структуры ядра нуклеосомы, полученной в 1997 году, удалось установить ключевые факторы формирования октамерного комплекса и взаимодействия его с ДНК, которая организована в суперспираль вокруг него [3]. Ядро нуклеосомы образуется из пары димеров H2A-H2B и тетрамера H3-H4. Основными участниками сцепления в гистон-гистоновых и гистон-ДНК взаимодействиях являются глобулярные домены гистонов, именуемые “histone fold” [3]. Они структурно схожи для всех четырех типов гистонов и состоят из трех спиралей, соединенных двумя петлями: L1 и L2. В то же время у гистонового семейства H2A наблюдается расширение данного домена небольшой αC-спиралью. αC-спираль и C-концевой хвост H2A образуют домен, именуемый "docking domain", который фиксирует димер H2A-H2B на поверхности тетрамера H3-H4, образуя полное ядро нуклеосомы [3,17]. Октамерный комплекс гистонов, вокруг которого оборачивается примерно 145-147 п.о. хромосомной ДНК, встречается практически каждые 200±40 п.н. во всех эукариотических геномах [3]. Важно отметить, что такие структурные единицы гистонов как N- и C- концевые хвосты выступают наружу над изгибами суперспирали ДНК и между ними, чтобы контактировать с соседними частицами [3]. Хвосты гистонов подвержены различным посттрансляционным модификациям, которые способствуют разворачиванию ДНК, и могут взаимодействовать с транскрипционными факторами, которые меняют динамику хроматина [18]. Иными словами, хвосты гистонов позволяют регулировать экспрессию генов. Каждая нуклеосома в свою очередь стабилизируется линкерным гистоном H1. В структуре нуклеосомы H2A образует димеры с H2B посредством мотива «рукопожатия» (рис. 3).

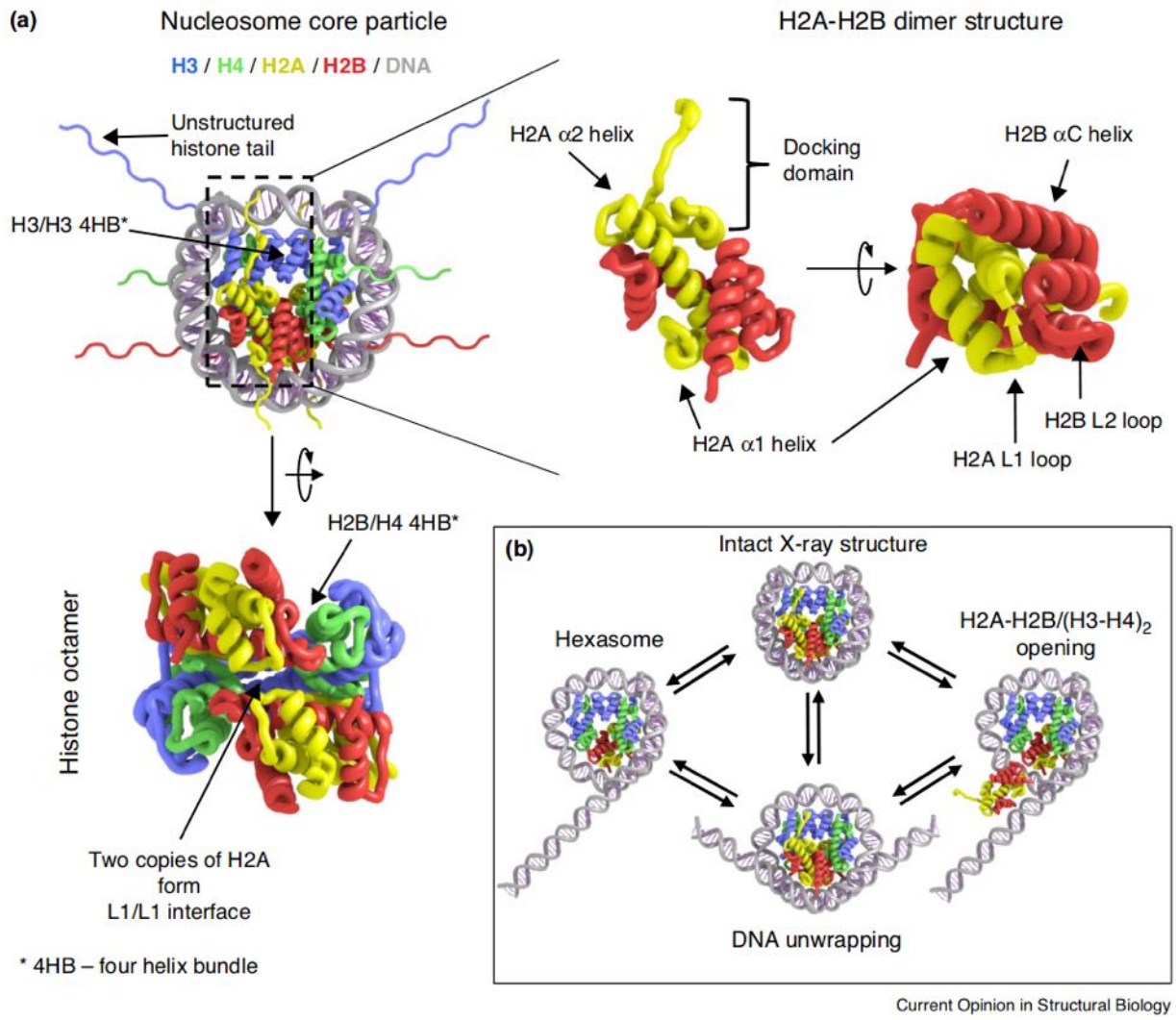

Рисунок 3. Структура и динамика нуклеосом со стороны димера H2A–H2B

Аминокислотные последовательности канонических гистонов являются высоко консервативными внутри каждого из 5 типов даже у отдаленно родственных видов, что демонстрирует их сильную эволюционную связь. Скорее всего это обусловлено тем, что их основной ролью является структурная упаковка ДНК, полученной в ходе репликации. Несмотря на высокое сходство существуют доказательства того, что незначительные различия между ними могут иметь функциональные последствия (например, влиять на стабильность нуклеосом).

Ожидаемо, что наиболее исследованными являются канонические гистоны у человека, у которого всего обнаружено 11 значимых изоформ из семейства H2A, 15 из семейства H2B, 1 из семейства H3 и 2 из семейства H4.

Для некоторых из них известны функциональные особенности. Например, у генов первого кластера на хромосоме 1 в положении 51 обнаружен метионин вместо лейцина, что приводит к изменению подвижности при электрофорезе мочевины в тритоновой кислоте [19]. А гены H2AC1 и H2BC1, кодирующие канонические гистоны H2A и H2B соответственно, считаются факторами материнского эффекта [20]. Как показано в одном из исследований, экспрессия этих генов усиливает перепрограммирование клеток, вызванное факторами Яманаки (OSKM-факторами), в клетках человека [16]. Более того, гены H2AC1 и H2BC1 являются самым дивергентным в своих семействах. Последовательности гистоновых белков cH2A.1 и cH2B.1, являющиеся их продуктами, на одну аминокислоту длиннее, чем у остальных представителей семейства. Известно также, что cH2A.1 отличается несколькими позициями внутри последовательности, а также в последних шести аминокислотах на С-конце по отношению к большинству канонических H2A. А cH2B.1 отличается от остальных канонических H2B почти на 25 % [21].

Наиболее сходными с человеческими считаются канонические гистоны мышей. В настоящее время в геноме мыши аннотировано 18 репликативно-зависимых генов H2A: 13 - в кластере на хромосоме 13, 4 - в кластере на хромосоме 3 и 1 - на хромосоме 11. Самый большой кластер содержит ген H2ac1, родственный человеческому H2AC1. Гистоновые белки, являющиеся продуктами генов H2ac1 и H2bc1 (кодирующие канонические гистоны H2A и H2B соответственно), так же как и человеческие, обнаружены в процессах перепрограммирования клеток [22]. Они способствуют активации отцовского генома после оплодотворения и “помогают” факторам Яманаки (OSKM-факторам) [22]. Известно также, что данная изоформа имеет меньше контактов с ДНК и нарушенные взаимодействия L1-L1-петли в рентгеновской структуре нуклеосомы. А на основе мутационного анализа in vivo [12] было выдвинуто предположение, что гистоновые хвосты и петля L1 важны для перепрограммирования. Также, в анализе дифференциальной сканирующей калориметрии было показано, что комплекс H2ac1/H2bc1 более

стабилен, чем другие комбинации канонических гистонов [22]. Несмотря на это, продукт мышиного гена H2ac1 отличается от своего человеческого паралога в 18 позициях.

Канонические формы H2A растений на данный момент изучены очень слабо [23]. У хлорофитных зеленых водорослей, таких как Chlamydomonas, канонические гистоны похожи на cH2A животных: кодирующие их гены сгруппированы в кластеры, а репликативно-зависимые мРНК заканчиваются петлей на 3'-стебле, подобно генам гистонов у животных [12,23]. В то же время у наземных растений наблюдается другая картина. Их мРНК полиаденилированы, а гены неупорядочены и распределены по всему геному. Более того, гены канонических гистонов наземных растений имеют интроны, что в принципе не свойственно репликативно-зависимым генам гистонов.

Интересно отметить, что у большинства видов грибов отсутствует каноническая форма H2A. Функции этого репликативно-зависимого белка у них выполняются вариантным гистоном H2A.X [24]. Согласно другим исследованиям, у некоторых паразитических форм грибов обратная ситуация. У них отсутствует гистоновый вариант H2A.X, а роль этого варианта берет на себя каноническая форма H2A [25].

Кроме того, единственный ген человеческого гистона H4, который расположен вне любого кластера гистоновых генов, экспрессирует неполиаденилированную мРНК в S фазе. Функция этого гена неизвестна; однако у мыши есть синтетическая копия, что позволяет предположить, что она сохранилась в эволюции млекопитающих [12].

Варианты

Динамика хроматина зависит от последовательности ДНК, ассоциации с различными транскрипционными факторами (например, FACT), гистоновых вариантов, составляющих ядро нуклеосомы, а также посттрансляционных модификаций, такие как ацетилирование, метилирование, фосфорилирование, убиквитинирование и сумоилирование [18]. Важную роль в динамике хроматина и регуляции экспрессии генов играют варианты

гистонов. Они могут встраиваться на место канонических гистонов в ходе всей жизнедеятельности эукариотической клетки, что приводит к различным эпигенетическим изменениям. Причем, особое место занимают хвосты гистоновых вариантов. Благодаря своему расположению (вне ядра нуклеосомы) они чаще всего подвергаются различным посттрансляционным модификациям и взаимодействию с транскрипционными факторами. Перечисленные факторы влияют на динамику развертывания нуклеосом и подвижность линкерной ДНК [18].

В результате исследования взаимосвязи между составом хроматина и его динамикой, были установлены режимы изменения компактности нуклеосом, а также факторы, влияющие на их динамику [18]. На рисунке 4а можно видеть следующие моды динамики нуклеосом:

1) скольжение - перемещение октамера гистонов вдоль ДНК в сочетании с изменением его вращательной установки;
2) подвижность линкерной ДНК;
3) внутренняя пластичность октамера гистонов;
4) спонтанное образование петель в нуклеосомной ДНК;
5) случайное разворачивание нуклеосомной ДНК;
6) образование нуклеосомных интермедиатов: гексасом, тетрасом, гемисом.

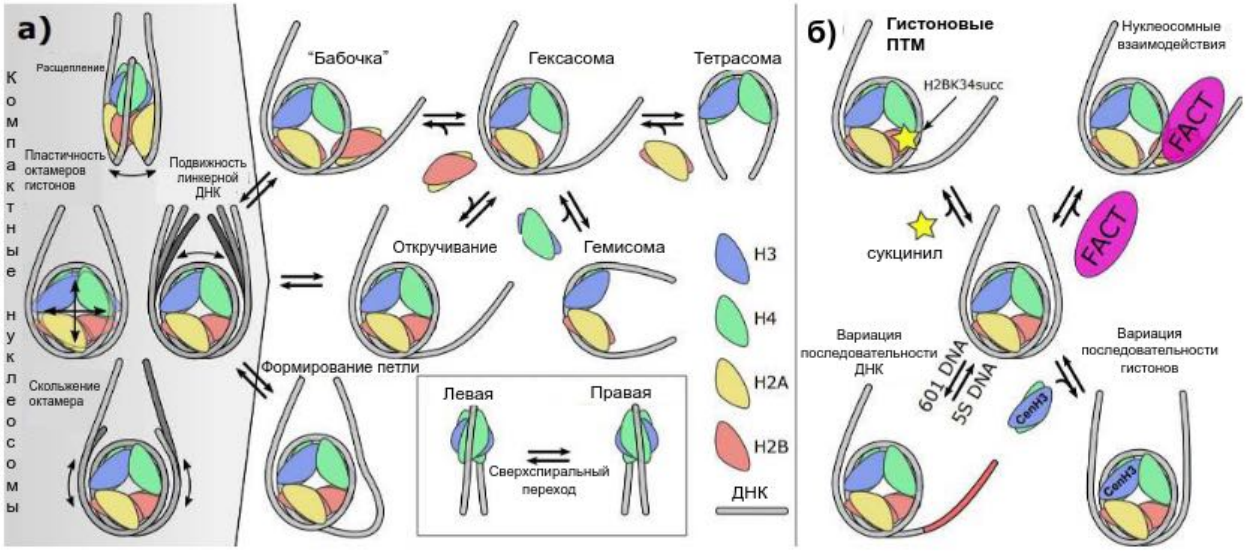

Рисунок 4. Концептуальная схема динамики структуры нуклеосом и факторов, на нее влияющих [18,26]. (а) Динамические режимы нуклеосом. Слева (серая область) проиллюстрированы режимы с сохранением компактности нуклеосом. Справа представлены режимы с большей амплитудой. (б) Примеры ключевых факторов, влияющих на динамику нуклеосом.

Среди ключевых факторов, влияющих на динамику нуклеосом, выделяют различные вариации последовательностей гистоновых белков, вариации последовательности ДНК, посттрансляционные модификации гистонов, а также разнообразные белки, взаимодействующие с нуклеосомами (рис. 4б) [18,26].

Гены гистоновых вариантов, в отличие от канонических изоформ, обычно кодируют полиаденилированные мРНК. Однако, интересно отметить вариант H2A.X, который участвует в распознавании повреждений ДНК в хроматине. Ген H2A.X кодирует как полиаденилированные, так и неполиаденилированные мРНК. Во время S-фазы клеточного цикла экспрессируется мРНК, 3'-конец которой не полиаденилирован. Но в фазах G0 и G1 мРНК H2A.X заканчивается поли(A)-хвостом [12]. Иными словами, гистоновый белок H2A.X синтезируется на протяжении всего клеточного цикла, что гарантирует конститутивную экспрессию мРНК H2A.X. Исследователи Marzluff William F. и др. предполагают, что на обеспечение

экспрессии разных форм мРНК из одного и того же гена повлияло то, что ген варианта H2A.X расположен вне основного кластера [12].

Наиболее исследованным семейством является H2A. Именно для этого типа гистонов описано наибольшее количество функциональных вариантов. Среди гистоновых вариантов H2A сегодня известны H2A.Z, H2A.X, macroH2A, short H2A, H2A.R, H2A.J, H2A.W, H2A.M, gH2A. Каждый из них экспрессируется независимо от репликации и играет важную функциональную роль в жизнедеятельности эукариотической клетки. Различия между вариантными гистонами присутствуют как на уровне аминокислотной последовательности, так и на структурном и функциональном уровнях.

Известно, что некоторые варианты отличаются от своей канонической формы лишь несколькими аминокислотными остатками, при этом сохраняется их структурная и функциональная роли. Например, варианты H2A.Z и H2A.X являются самыми ближайшими родственниками к cH2A [27]. H2A.Z отличается от cH2A и H2A.X в тех аминокислотных остатках, где соединяются петля L1 с α-спиралью α2 и α-спираль α2 с петлей L2, а также по C-концевому «стыковочному домену», который контактирует с H3 [5]. Также, дерево, полученное методом присоединения соседей, демонстрирует, что H2A.Z вероятнее всего отделились от других H2A до диверсификации современных эукариот, а H2A.X и cH2A, по-видимому, неоднократно расходились в разных линиях [5,27].

Так же как и канонические H2A, варианты близкие к нему, демонстрируют высокую консервативность внутри подсемейства. Например, H2A.Z строго консервативен почти у всех эукариот (исключения могут включать некоторые метамонады (например, Giardia, Trichomonas), амебы и паразитические грибы [25]). У хордовых известно два гена H2AZ1 и H2AZ2, продукты которых отличаются всего тремя аминокислотными остатками. Однако, важно отметить, что они приобрели некоторую степень функциональной независимости. Например, было показано, что H2A.Z.1

лучше взаимодействует с белком BRD2, содержащим бромдомен [28], а H2A.Z.2 предпочтительно связывается с H3, триметилированным по лизину 4 (H3K4me3) [29].

Следует подчеркнуть, что у близких вариантов также наблюдаются схожие функции. Например, варианты H2A.Z и H2A.X участвуют в процессах репарации ДНК [30]. При этом наличие специфических особенностей на уровне аминокислотной последовательности влияет на наличие и специфических функций. Например, у H2A.Z область петли L1 отличается четырьмя аминокислотами от cH2A и, вероятно, участвует в придании стабильности и функциональной специфичности вариантным нуклеосомам посредством взаимодействий L1-L1 [17]. А С-концевая область дрожжевого белка H2A.Z взаимодействует с РНК-полимеразой II (RNAPII) и тем самым способствует ее рекрутированию на промоторы [30,31]. Вариант H2A.X имеет характерный мотив SQE/DФ на С-конце (SQAY у дрозофилы), где Ф-представляет собой гидрофобный остаток (обычно Tyr у млекопитающих). Остаток серина (S) представляет собой сайт фосфорилирования. С помощью своей фосфорилированной формы ɣ-H2A.X он способен отмечать разрывы двухцепочечной ДНК и тем самым реагировать на повреждение ДНК [32].

Другие же варианты могут отличаться более, чем на 25% не только от канонической формы, но и внутри подсемейства. Самую низкую идентичность с cH2A демонстрируют короткие (short) H2A. Например, известно, что аминокислотная последовательность H2A.B всего на 50% идентична cH2A, H2A.L - на 30%, а H2A.P - всего на 24% [33].

Структурные и функциональные признаки гистоновых вариантов могут быть очень разнообразными. Например, белки подсемейства short H2A имеют укороченные С-концевой хвост и домен, известный как “docking domain” (свойственный всем H2A), а также отсутствие некоторых кислотных остатков в регионе кислотного лоскута, именуемого “acidic patch”, который отвечает за стабильность в нуклеосомно-нуклеосомных взаимодействиях. Ввиду этого

нуклеосомы, включающие короткие H2A, обертывают меньшую ДНК (120-130 п.н.) и образуют рыхло упакованный хроматин [33].

Некоторые представители H2A имеют специфичные домены или мотивы. Помимо H2A.X, который характеризуется мотивом SQE/DΦ на С-конце, можно отметить вариант H2A.W. Он имеет мотив SPKK (иногда называемый KSPKKA) в своем С-концевом хвосте, потенциально связывающийся с малой бороздкой ДНК [34]. Известно, что H2A.W чаще всего встречается в гетерохроматине, участвует в подавлении активности генов, а также реагирует на повреждение ДНК [7,35]. Другим интересным примером является macroH2A, который отличается наличием длинного негистонового макродомена (~30 кДа), именуемого “macro domain”, соединенным с С-концом глобулярного домена через неструктурированный линкер. Изучение функциональных особенностей macroH2A привело к выводу, что макродомены могут быть способны связывать производные метаболиты NAD+ , такие как АДФ-рибоза и поли-АДФ-рибоза [36].

Гистоновые варианты могут быть специфичны к отдельным таксономическим группам или клеточным культурам. Например, варианты H2A.W, H2A.M и gH2A обнаружены исключительно в растениях. Известно, что H2A.W и H2A.M - родственные гистоны, однако, первый из них найден только в покрытосеменных (цветковых) растениях, а второй - в нецветковых растениях, таких как печеночники, мхи и ликофиты. А также H2A.M отличается от других H2A наличием длинного С-концевого хвостового домена, богатого лизином, серином и кислотными остатками [7]. В отличие от H2A.W и H2A.M вариант gH2A, обнаруженный в роде Lilium, исследован очень слабо. На данный момент он считается специфичным для мужской гаметы [23,37]. Еще одним примером являются короткие гистоновые варианты H2A. Они представляет собой класс, охватывающий несколько вариантов гистонов семейства H2A у плацентарных млекопитающих, экспрессирующихся в основном во время развития мужских половых клеток

млекопитающих до почти полной замены гистонов протаминами в ядрах сперматозоидов [33].

Различия между гистоновыми вариантами можно также отметить на транскрипционном уровне. Бóльшая часть генов, кодирующих варианты семейства H2A, имеет интроны. Однако существуют гистоны, кодирующие гены которых не содержат интронов.

1.3 Базы данных гистоновых белков

Первая база данных гистоновых вариантов появилась в 1995 году. Она была разработана исследователями из National Center for Biotechnology Information (NCBI) и получила название Histone Sequence Database и хранила более 1300 аминокислотных и нуклеотидных последовательностей гистоновых белков. Для наполнения данной базы данных использовался поиск с помощью программы BLASTP на последовательностях из SwissProt 31.0, PIR 45.0, GenPept 91.0 и PDB. Для создания базы данных (BLASTDB), относительно которой осуществлялся поиск, строились выравнивания аминокислотных последовательностей гистоновых белков человека и курицы. При этом аминокислоты histone fold, структурного мотива характерного весем гистоновым белкам, выделялись в отдельную рамку [38].

Следующие два года Histone Sequence Database успела претерпеть множество изменений. Важной особенностью обновления стало применение двух алгоритмов для поиска гистоновых белков: BLASTP и PSI-BLAST. При этом для построения парных и множественных выравниваний использовался инструмент CLUSTALW, а для обнаружения глобулярного домена - программы Motif Search Tool и PROBE (Baxevanis, Landsman, 1998). Чуть позже в базу данных была добавлена информация о посттрансляционных модификациях, локусах генов, а также данные геномных исследований. Более того, чтобы избавиться от “избыточной” (redundant) информации, в Histone Sequence Database были выделены две отдельные части. В первую поместили полный набор, в котором перечислены все последовательности, найденные для этого типа гистона. А вторая содержала информацию о гистоновых

вариантах, последовательность которых была определена лишь один раз (non-redundant) (Makalowska et al., 1999).

Спустя год база данных была переименована в The Histone Database (или HDB), а также была опубликована статья с презентацией ее очередного обновления. HDB была разделена на 10 частей [39]:

1) исходные данные и общая информация по ним, включая источники;
2) поисковая система, параметры которой включают тип белка, набор последовательностей, микроорганизм, ключевое слово или шаблон последовательности;
3) все аминокислотные последовательности эукариотических гистоновых белков в формате FASTA;
4) “неизбыточный” набор одинаковых последовательностей в формате FASTA;
5) все аминокислотные последовательности архейных гистонов и негистоновых белков, содержащие глобулярный домен, в формате FASTA;
6) множественные выравнивания всех белков последовательностей, включая выравнивания глобулярных доменов гистонов архей и негистоновых белков, выполненные с использованием CLUSTALW;
7) таблицы 3D-структур гистоновых белков, включая глобулярные домены, доступные в базах данных трехмерных структур;
8) сводка посттрансляционных модификаций гистоновых белков;
9) графическое изображение семи хромосом человека (I, IV, VII, XI, XVII, XXII) с выделением локусов генов гистонов;
10) список несоответствий между записями об одной и той же последовательности в первичных базах данных, таких как GenBank.

В последующие 10 лет база данных HDB продолжала обновляться и к 2006 году общее число последовательностей превысило 3000, а количество видов живых организмов достигло 975 [40,41]. За это время обновились также используемые алгоритмы множественного выравнивания [41].

В 2011 году впервые коллекция базы данных была расширена путем поиска гомологичных последовательностей с использованием программы HMMER3 [42]. В базу данных также были добавлены последовательности из нового источника Protein Research Foundation (PRF). Количество таксономических идентификаторов составило более 7300, что примерно соответствовало количеству организмов [42].

Наконец, в 2015 году вышла еще одна версия базы данных гистоновых белков, которая получила название HistoneDB 2.0 [43]. Было произведено самое масштабное обновление, в результате которого база данных включила более 80 тысяч аминокислотных последовательностей гистоновых белков, классифицированных по 30 различным семействам. Одним из важных нововведений было использование новой филогенетической номенклатуры гистонов, предложенной в 2012 году [6]. HistoneDB 2.0 состоит из двух наборов данных. Один из них представляет собой курируемые аминокислотные последовательности гистонов со ссылками на литературные источники. Второй набор сгенерирован автоматически с помощью алгоритмов поиска гомологичных последовательностей. Клиентское приложение доступно по ссылке https://www.ncbi.nlm.nih.gov/research/HistoneDB2.0/ и позволяет не только осуществлять поиск, но и строить выравнивания и филогенетические деревья, что позволяет анализировать последовательности, хранящиеся в базе данных. Кроме того, пользователь может исследовать свою последовательность. Для этого применяются встроенные алгоритмы классификации гистоновых белков на различные семейства, которые также использовались для генерации автоматически полученного набора последовательностей базы данных [43].

Глава 2. Методы анализа разнообразия гистонов

2.1 Биоинформатические методы анализа аминокислотных последовательностей

Для написания скриптов в ходе проведения анализа разнообразия и классификации гистоновых белков были использованы прикладные библиотеки высокоуровневого языка программирования Python 3, в число которых вошли biopython, pytexshade, pynucl, matplotlib, seaborn, scipy, numpy, pandas, ete3, Django и др.

Python - это интерпретируемый язык со строгой динамической типизацией и автоматическим управлением памятью, впервые возникший в 1980-х годах. Основными преимуществами Python являются качество программного обеспечения, высокая скорость разработки, переносимость программ, библиотеки поддержки и интеграция компонентов. Единообразие оформления программного кода на языке Python повышают его читаемость и облегчают понимание не только для его разработчиков, но и для лиц, не участвовавших в написании кода. Так как Python интерпретируемый язык, сокращается время на отладку и требуется меньший объем трудозатрат на сопровождение. Программы, написанные на языке Python легко переносить. Для этого обычно достаточно простого копирования файлов программ с одной машины на другую.

Важно отметить, что Python допускает расширение за счет различных сторонних библиотек. Имеется возможность подключения библиотек для всевозможных математических и статистических вычислений, для разработки алгоритмов на основе машинного обучения, а также поддерживающие различные инструменты, специфичные некоторой области. Например, Biopython, созданная международной ассоциацией разработчиков, представляет из себя коллекцию некоммерческих инструментов для вычислительной биологии и биоинформатики. Также в числе сторонних библиотек есть инструменты для создания веб-сайтов, что является

актуальной задачей при разработке и обновлении базы данных Histone Database.

В настоящее время существует две основные версии: Python 2, вышедшая в 2000 году, и Python 3, выпущенная в 2008 году и в настоящее время находящаяся в разработке. Версия Python 3 стремится учесть и исправить структурные недостатки предыдущих релизов. Однако, она не совместима с Python 2, так как имеет широкий ряд важных отличий. Ввиду того, что поддержку Python 2 прекратили с 2020 года, был осуществлен рефакторинг кода ранее разработанной базы данных Histone Database для совместимости с Python 3. А также для дальнейшей работы с обновлениями и анализом аминокислотных последовательностей гистоновых белков использовалась одна из последних версий - Python 3.7.

Для проведения комплексного анализа разнообразия аминокислотных последовательностей гистоновых белков в широком спектре живых организмов были построены глобальные и множественные выравнивания.

Выравнивания является одним из важнейших инструментов при сравнении биологических последовательностей. Они позволяют обнаружить схожие участки и выявить функциональные, структурные или эволюционные взаимосвязи между первичными структурами. В зависимости от количества анализируемых последовательностей выравнивания можно разделить на 2 типа: парные и множественные. Чтобы оценить сходства двух последовательностей обычно строятся парные (Pairwise alignment) выравнивания. При этом, если предполагается, что последовательности гомологичны по всей длине, используются алгоритмы глобального выравнивания, например, Нидлмана-Вунша (https://doi.org/10.1016/0022-2836(70)90057-4). В случае необходимости обнаружить только отдельные гомологичные участки двух последовательностей применяется применяются алгоритмы локального выравнивания, например, Смита-Ватермана

(https://.doi.org/10.1016/0022-2836(81)90087-5). Оба алгоритма основаны на принципах динамического программирования.

В данной работе были использованы алгоритмы Нидлмана-Вунша и множественные выравнивания, так как предполагается, что аминокислотные последовательности гистоновых белков гомологичны по всей длине, а также они эволюционно связаны. Для этого применялись готовые программы, реализующие необходимые алгоритмы: EMBOSS Needle для построения парных глобальных выравниваний, MUSCLE и CLUSTALW2 для построения множественных выравниваний.

Для проведения эволюционного анализа с целью сократить влияние разнообразия неупорядоченных хвостов гистонов были извлечены центральные области глобулярных доменов гистонов, гистоновые складки, (histone fold domain, HFD), являющиеся структурными мотивами, характерными для всех гистонов, состоящими из трех альфа-спиралей, соединенных двумя петлями. Множественные выравнивания полученных последовательностей строились с применением программы MUSCLE [17]. Филогенетические деревья были построены с помощью алгоритмов PhyML [18], в основе которых лежат методы максимального правдоподобия.

На основе множественных выравниваний был проведен эволюционный анализ. Филогенетические деревья построены в программном обеспечении PhyML с использованием алгоритмов Neighbor-Joining и поиска максимального правдоподобия. Также, чтобы обнаружить новые значимые подсемейства гистоновых белков, была проведена кластеризация на основе метода UPGMA.

Для обновления и улучшения алгоритмов поиска и классификации аминокислотных последовательностей гистоновых белков была разработана модель, позволяющая отбирать гистоны из обширного множества других белков, а также определять семейство гистонового варианта. Ее основными компонентами являются программы для поиска гомологичных белков, такие как BLASTP и HMMER.

Все алгоритмы для поиска гомологичных последовательностей опираются на множество локальных выравниваний. Общая схема заключается в том, чтобы сравнить исследуемую последовательность с некоторой базой данных последовательностей и найти список наиболее значимых выравниваний с высоким сходством.

Одним из известных инструментов является программный пакет HMMER, написанный Шоном Эдди. Его методы основываются на построении скрытой марковской модели (profile HMM), которая представляет из себя вероятностную модель, имитирующую работу процесса, похожего на марковский процесс с неизвестными параметрами. Она используется в задачах, где необходимо предсказать неизвестные параметры на основе наблюдаемых. Сравнивая с ней исследуемые последовательности, можно сделать вывод о гомологии.

Алгоритм поиска гомологичных последовательностей с помощью HMMER:

1. Построение множественного выравнивания известных гомологичных последовательностей.
2. Построение скрытой марковской модели, опираясь на полученное выравнивание.
3. Поиск последовательностей в исследуемой базе данных путем сравнения со скрытой марковской моделью.

Для каждой обнаруженной в базе данных последовательности HMMER рассчитывает не только вес, отражающий сходство с исходной последовательностью, но и то, насколько значимо само сходство. Эта величина называется значимостью выравнивания, или expected value, E-value. E-value - это количество ожидаемых выравниваний аналогичного качества (оценка), которые могут быть обнаружены случайно. Например, E-value, равное 10, означает, что можно ожидать до 10 выравниваний, полученных случайным образом, учитывая тот же размер случайной базы данных. Чем меньше E-value, тем более значимо выравнивание.

Для оценки метрики близости HMMER использует несколько параметров, характеризующих степень гомологии обнаруженного участка последовательности:

- E-value позволяет отобрать наиболее значимые последовательности;
- Score - оценка близости последовательности из исследуемой базы данных к скрытой марковской модели;
- Bit Score - score с учетом смещения.

Другим важным инструментом является семейство компьютерных программ BLAST (англ. Basic Local Alignment Search Tool - средство поиска основного локального выравнивания), предназначенное для поиска гомологов белков или нуклеиновых кислот, для которых известна последовательность или ее фрагмент. Основной целью метода является поиск всевозможных локальных выравниваний (High Scoring Pairs, HSPs) по исследуемой базе данных, удовлетворяющих критериям поиска.

Алгоритм поиска гомологичных последовательностей с помощью BLAST:

1. Для каждого слова длины W в исходной последовательности производится поиск "похожих" слов таких, что вес их выравнивания выше некоторого порога T. В большинстве программ серии BLAST используется матрица BLOSUM62 (Blocks Substitution matrix 62 % identity) (https://doi.org/10.1073%2Fpnas.89.22.10915). Исключением являются blastn и megablast (программы, которые выполняют нуклеотид - нуклеотидные сравнения и не используют матрицы аминокислотных замен).
2. Производится поиск "похожих" слов в исследуемой базе данных, т.е. для каждого слова находятся все вхождения в последовательности, хранящиеся в БД. Этот поиск производится на основе составленной предварительно "библиотеки слов" - хэш‑таблице.
3. Выравнивания длины W расширяются вправо и влево с использованием алгоритма динамического программирования. Эта процедура

происходит до тех пор, пока падение суммарного веса выравнивания от точки последнего максимума не достигнет некоторого порога X. Длина выравнивания HSP будет равна значению, соответствующему позиции последнего максимума.

Для оценки метрики близости BLAST использует несколько параметров, характеризующих степень гомологии обнаруженного участка последовательности:

- E-value позволяет отобрать наиболее значимые последовательности;
- Score - оценка близости последовательности;
- Bit Score - score с учетом смещения.

2.2 Разработка веб-сервиса и базы данных

Для того чтобы провести техническую модернизацию базы данных Histone Database были использованы такие инструменты, как Django Framework и MySQL.

В качестве система управления базой данных была выбрана свободная реляционная СУБД MySQL, разработанная и поддерживаемая корпорацией Oracle. Ее основными преимуществами являются: наличие API [https://dev.mysql.com/doc/refman/8.0/en/connectors-apis.html] и коннекторы для языков для множества разных языков, таких как Delphi, C, C++, Java, Python и др., а также поддержка большого количества типов таблиц.

Для реализации веб-приложения на языке Python был выбран свободный веб-фреймворк Django, поддерживаемой организацией Django Software Foundation. В его основе лежит использование концепция проектирования Model-View-Controller (MVC), предполагающая разделение данных приложения на 3 независимых компоненты:

- модель (Model), которая является объектом для хранения данных и взаимодействует с контроллером для модификации своего состояния;
- представление (View), позволяющее представлять объект данных пользователю, а также возможность отвечать на изменения модели;

- контроллер (Controller), который интерпретирует действия пользователю и отправляет запросы в модель для необходимости изменения ее состояния.

Более того в Django реализован собственный ORM для работы с базой данных, что значительно облегчает настройку взаимодействия веб-сервера с базой данных. При этом модель данных в Django ORM описывается объектами Python.

Для повышения устойчивости и переносимости приложения использовалось программное обеспечение Docker, которое позволяет автоматически развертывать систему в средах с поддержкой контейнеризации. Такой подход позволяет “заворачивать” программную систему вместе со всем его окружением и зависимостями в “обертку”, которую в дальнейшем можно перенести и развернуть на любой другой операционной системе с поддержкой контейнеризации. Обертка приложения называется образом, или Docker-образом, а развернутая из него система называется контейнером, или Docker-контейнером.

Глава 3. База данных HistoneDB 3.0

3.1 Содержание и структура базы данных

База данных HistoneDB 3.0 является модернизацией предыдущей версии HistoneDB 2.0. Ее основное содержание представлено на рис. 5.

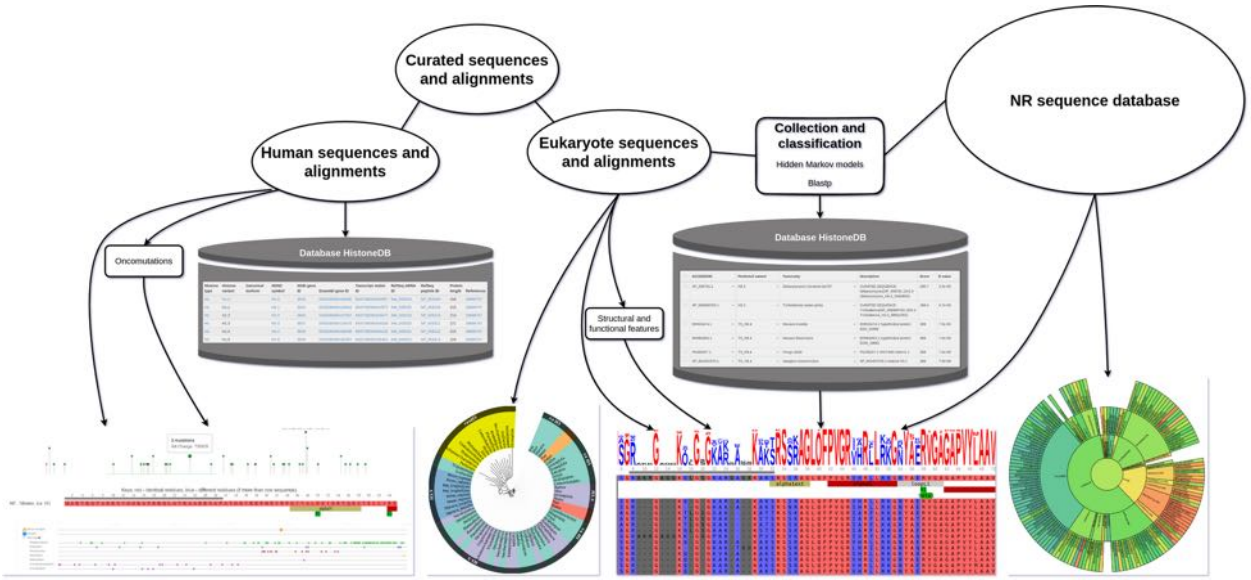

Рисунок 5. Схема основного содержания базы данных HistoneDB 3.0

HistoneDB 3.0 состоит из двух наборов данных: курируемый и извлеченный автоматически из баз данных, содержащих избыточную информацию о различных белковых последовательностях (рис. 5). Аминокислотные последовательности курируемого набора были собраны из различных литературных источников и проаннотированы вручную. Остальные последовательности были получены в результате отбора и классификации гистоновых последовательностей из базы данных NCBI (nr). Курируемый набор включает также коллекцию всех известных человеческих гистоновых белков, аннотация которых содержит название (HGNC symbol) и идентификатор (NCBI gene ID) гена, идентификаторы транскриптов и белков, а также длину белка и ссылки на источники.

Концептуальная модель базы данных состоит более, чем из 15 связанных таблиц (рис. 6 и 7). Ключевой таблицей django-приложения browse является Sequence. Каждая ее запись представляет собой аминокислотную последовательность с описанием, которое включает идентификатор

(ACCESSION VERSION), вторичный ключ, ссылающийся на запись из таблицы Variant, вторичный ключ, ссылающийся на запись из таблицы Taxonomy, а также другая актуальная информация о белке. Таблицы Type и Variant хранят информацию о типах и семействах гистоновых белков, соответственно. Они в свою очередь ссылаются на таблицу Publication, которая содержит записи об источниках, в которых описаны исследования с гистоновыми белками данного типа или семейства. Таблицы TemplateSequence и Features позволяют хранить информацию о структурных особенностях гистоновых семейств. Таблицы HmmScore и BlastScore содержат информацию обо всех значимых локальных выравниваниях, а также оценку их качества.

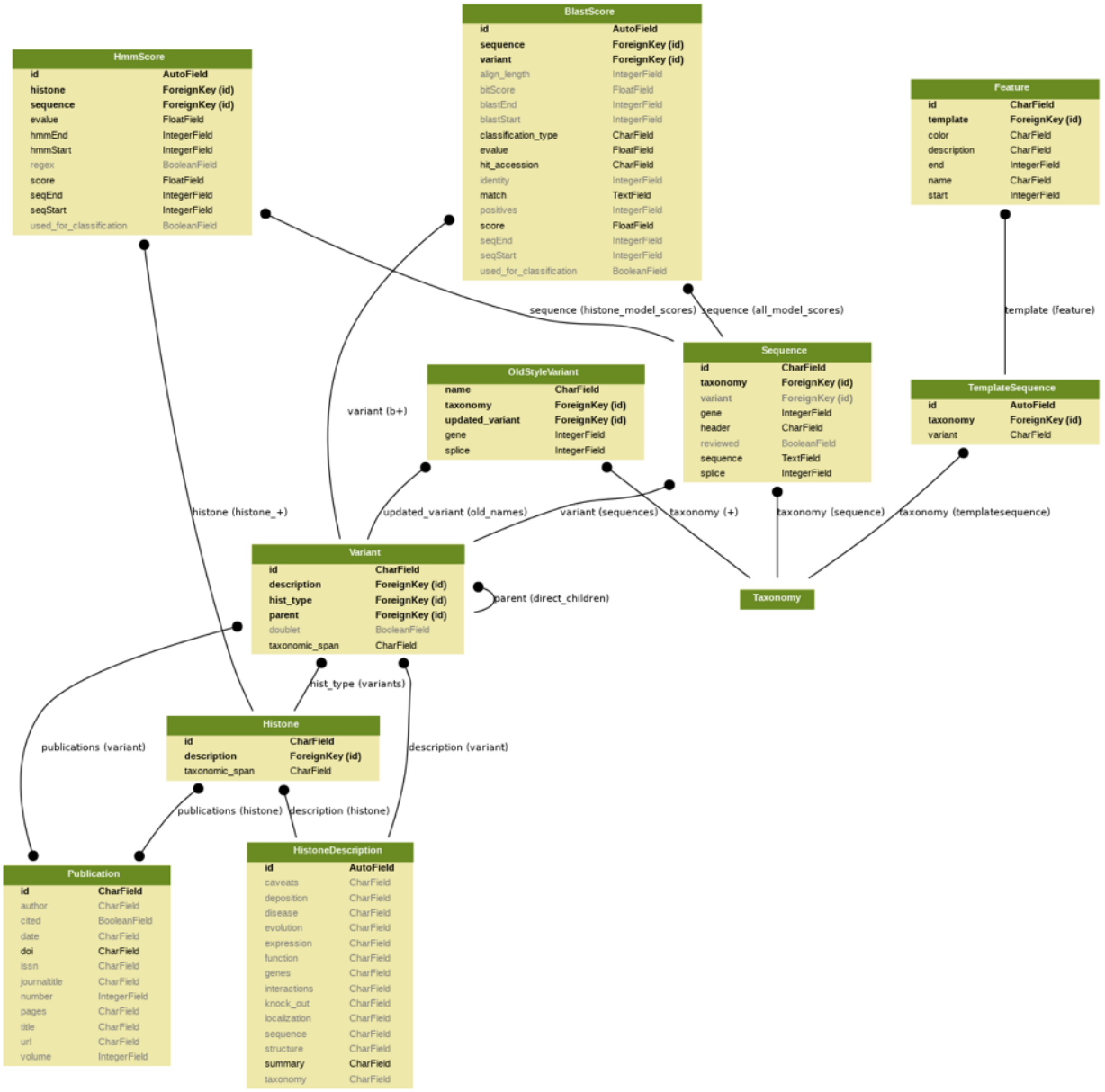

Рисунок 6. Концептуальная модель django-приложения browse базы данных HistoneDB 3.0

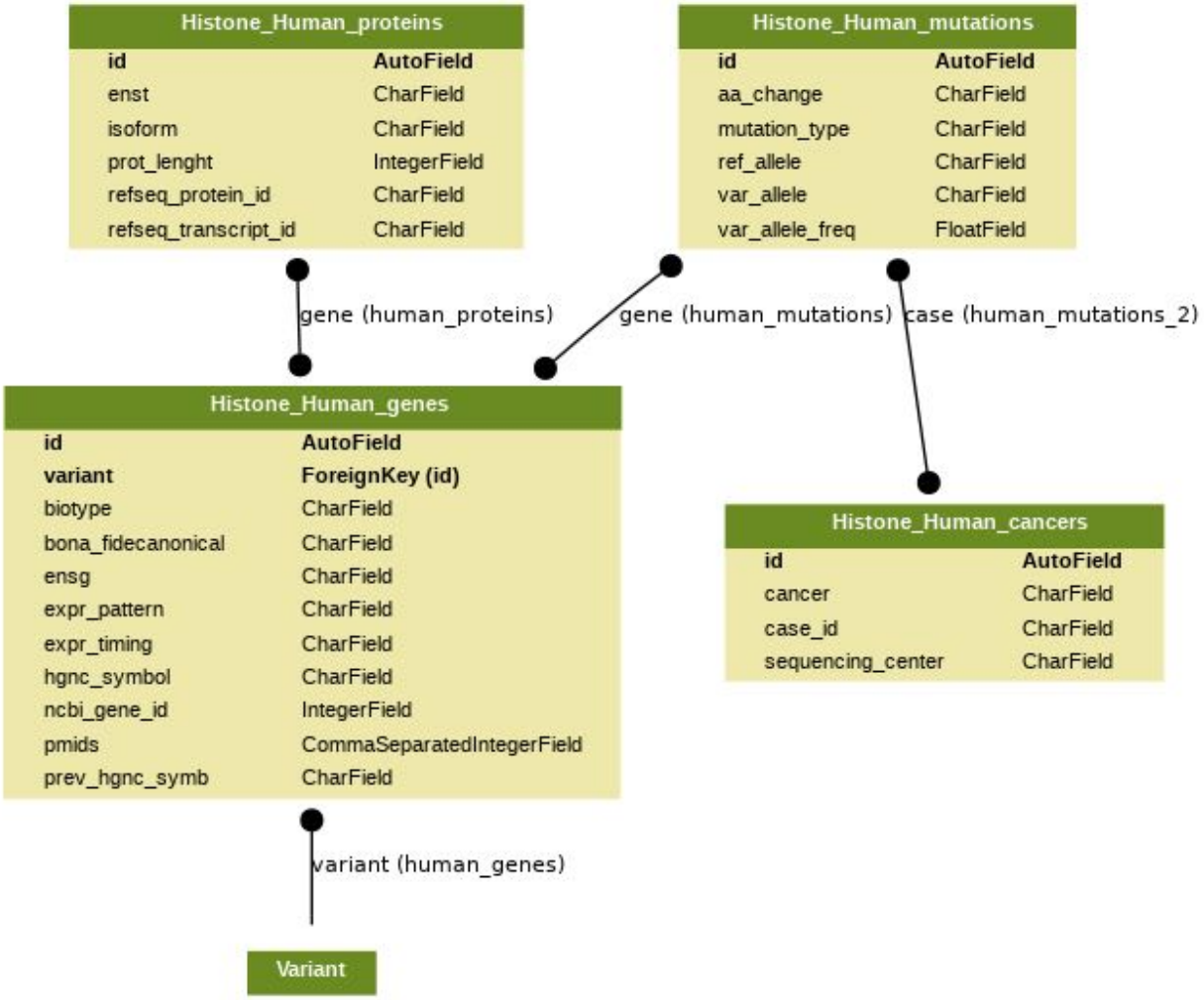

Рисунок 7. Концептуальная модель django-приложения human_hist базы данных HistoneDB 3.0

Еще одним нововведением является добавление более подробной информации обо всех известных человеческих гистонах. Для этого было разработано отдельное django-приложение human_hist, состоящее из четырех таблиц, которые хранят информацию о человеческих гистоновых белках и генах, кодирующих их. Таблица Histone_Human_proteins содержит идентификатор записи, название изоформы для канонических гистонов, идентификатор транскрипта, длину аминокислотной последовательности, а также вторичный ключ, ссылающийся на записи из таблицы Histone_Human_genes. Таблица Histone_Human_genes хранит информацию о всех генах гистонов человека, для которых определены название (HGNC symbol) и идентификатор (NCBI gene ID), ссылки на источники и другое. В

таблицах Histone_Human_mutations и Histone_Human_cancers хранится информация об онкомутациях в генах гистонов человека.

В ходе технического обновления базы данных HistoneDB произведен перенос системы на новый сервер. В связи с тем, что идентификатор GI устарел, произведена замена идентификатора записей на актуальный - ACCESSION VERSION. Также в целях переносимости программы был собран docker-образ, который позволяет разворачивать базу данных практически на любой системе.

3.2 Алгоритмы классификации

Разработка алгоритма классификации гистоновых белков на различные семейства является непростой задачей. Аминокислотная последовательность между вариантами гистонов может отличаться, как значительно, так и всего несколькими аминокислотами. Более того, некоторые семейства характеризуются специфическими мотивами или дополнительными доменами, играющими важную функциональную роль. Ввиду такого широкого разнообразия, для разработки моделей обнаружения и классификации гистонов в широком спектре живых организмов требуются алгоритмы, которые учитывают особенности их первичной структуры.

База данных HistoneDB 2.0 использует технологию HMMER, которая позволяет искать гомологичные последовательности на основе скрытых марковских моделей. Схема алгоритма представлена на рисунке 8 [44]. Для каждого варианта гистона отбираются курируемые аминокислотные последовательности для построения множественного выравнивания и обучения скрытой марковской модели, которая отражает статистические характеристики близости группы последовательностей. Из базы данных NCBI (nr) происходит отбор гистоновых последовательностей с помощью HMMsearch. Если значимость выравнивания более 10, такая последовательность считается негистоновой. Для классификации отобранных последовательностей по гистоновым вариантам используется оценка значения метрики близости (score). Если она выше порогового уровня,

вычисленного для каждого гистонового варианта заранее на основе курируемого набора, то последовательность считается принадлежащей данному варианту.

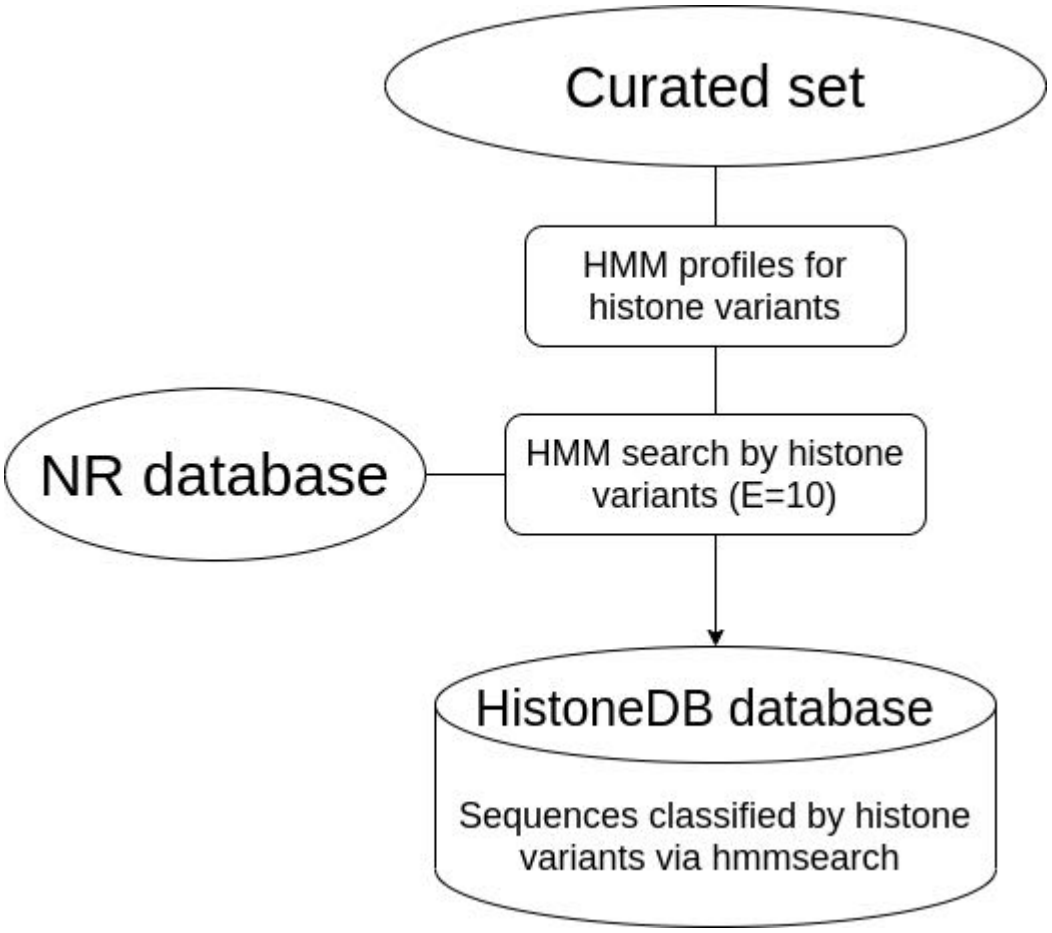

Рисунок 8. Схема алгоритма отбора и классификации гистоновых белков в базе данных HistoneDB 2.0

К сожалению, база данных HistoneDB 2.0 использует алгоритмы, которые позволяют лишь частично принять во внимание уникальные свойства гистоновых семейств. В ходе статистического анализа содержащихся в ней аминокислотных последовательностей были получены результаты, которые в целом совпадают известной нам информацией о гистоновых семействах из литературных источников. Однако, в некоторых случаях были обнаружены недостатки алгоритмов отбора и классификации гистонов.

Некоторые аминокислотные последовательности гистонов попали в семейства, которые специфичны другой таксономической группе. Например,

целый ряд гистоновых белков группы SAR (111 последовательностей) был отнесен к варианту H2A.W, который присутствует исключительно в растениях, как известно из актуальных исследований о нем. Аналогично, в семействе H1.0 были обнаружены последовательности грибов (296 последовательностей), у которых ранее не был засвидетельствован данный вариант. Приведенные примеры относятся к несовершенствам алгоритмов классификации, так как построенные множественные выравнивания и филогенетические деревья для данных последовательностей в объединении с курируемыми не продемонстрировали наличие каких-либо эволюционных взаимоотношений.

Кроме того, в результате статистического анализа было выявлено, что некоторые аминокислотные последовательности не прошли пороговые значения алгоритмов, то есть минимальную оценку метрики близости HMM. Например, в базу данных попало очень мало последовательностей варианта cenH3 растений (17 последовательностей), канонического H3 и H3.3 грибов (8 и 5 последовательностей, соответственно), а также канонического H2B группы SAR (4 последовательности). Опираясь на литературные источники, мы можем утверждать, что их должно быть больше.

Построенные множественные выравнивания некоторых последовательностей свидетельствуют о наличии ошибочно классифицированных гистонов. Например, последовательности с идентификаторами JQ1984, A5PK61.1, XP_005594567.1, XP_014983621.1 были отнесены к семейству H3.Y. Однако, у них отсутствует мотив, свойственный данному семейству, а также, опираясь на множественные выравнивания с другими семействами гистонов H3, они проявляют большую идентичность с курируемыми последовательностями семейства H3.3, нежели H3.Y.

Наряду с перечисленными недостатками можно отметить, что некоторые гистоновые варианты представлены очень низким количеством последовательностей (1-2 последовательности) в курируемом наборе. Из-за

недостатка примеров, скрытая марковская модель может быть слишком чувствительной к малозначимым особенностям, которые характерны всем первичным структурам белков гистонового семейства. Несмотря на это, важно отметить, что использование HMMER позволяет обнаружить дальних гомологов, что будет полезно для определения типа гистна.

Для того, чтобы устранить обнаруженные недостатки и обновить базу данных, алгоритмы отбора и классификации гистоновых белков были разделены на два основных этапа (рис. 9). Первый из них решает задачу отбора гистоновых аминокислотный последовательностей и “избыточной” базы данных белков NCBI (nr). Для этого используются скрытые марковские модели (HMM), которые обучены на тщательно отобранных выравниваниях глобулярных областей гистонов для каждого типа гистонов. После выполнения поиска гомологичных последовательностей для дальнейшей классификации остаются выравнивания с уровнем значимости ниже 10. Для определения типа гистона оценивается значение метрики близости (score) по максимальному значению. На следующем этапе отобранные последовательности группируются по типам гистонов и подразделяются на различные подсемейства. Для этого используется программа поиска гомологичных последовательностей BLAST. Ее преимущество в том, что путем построения выравниваний она позволяет вычислить идентичность последовательностей и обнаружить близкородственных гомологов. В самом конце этого этапа классификация уточняется путем рассмотрения мотивов, специфичных для вариантов, и доменов, специфичных для вариантов, в последовательностях. В приложении А можно ознакомиться с листингом кода, реализующего новый алгоритм поиска и классификации гистоновых белков.

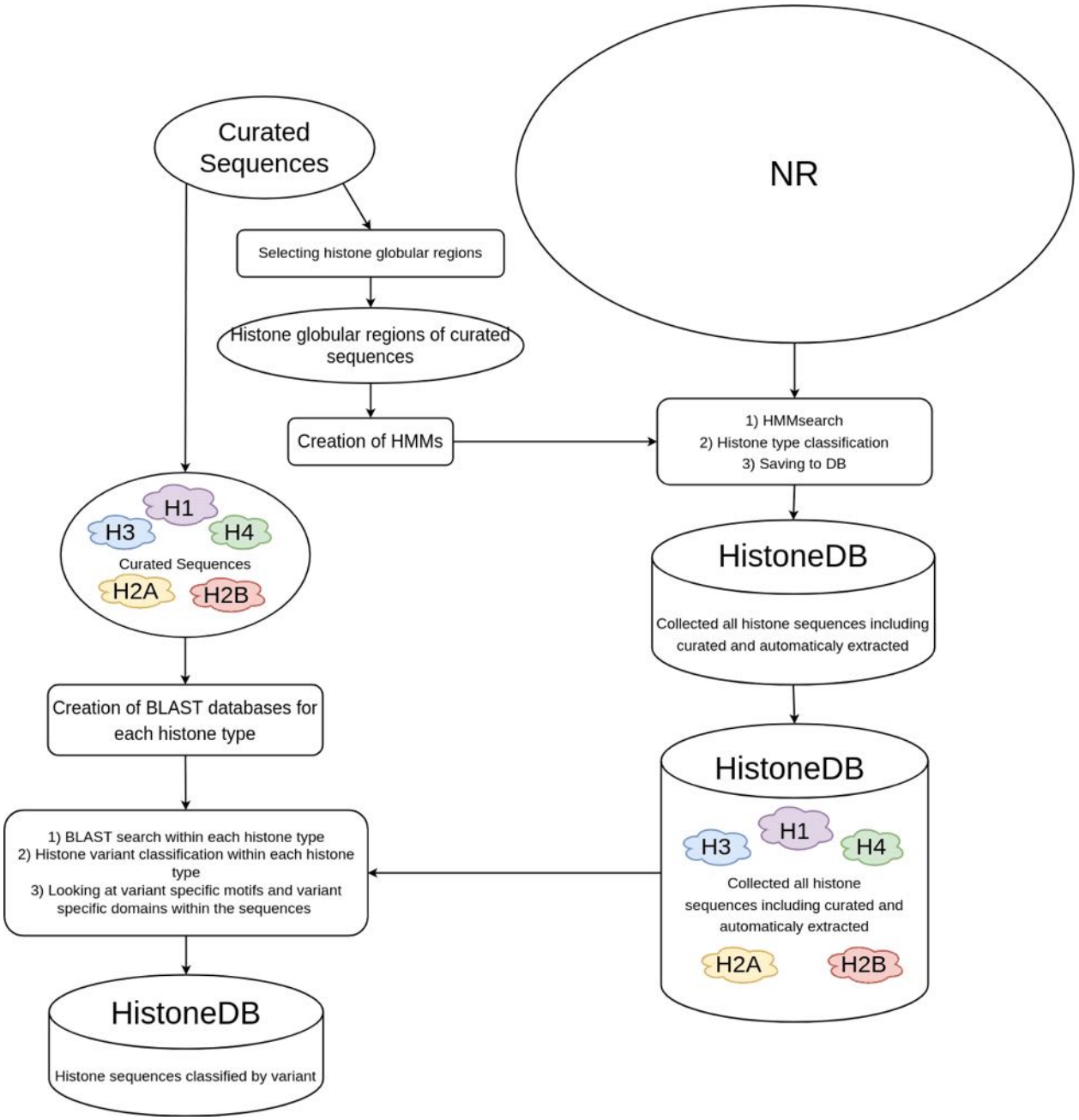

Рисунок 9. Схема алгоритма отбора и классификации гистоновых белков в базе данных HistoneDB 3.0

В результате, обновленная база данных гистонов HistoneDB 3.0 содержит более 186 тысяч аминокислотных последовательностей гистонов и более 34 различных подсемейств, а также обеспечивает более чувствительный алгоритм для обнаружения специфических особенностей вариантов. Также важно отметить, что новые алгоритмы справились с теми случаями, в которых ранее были допущены ошибки. Например, последовательностей, классифицированных как канонический H3 и H3.3 у грибов, теперь более

1500 и 13000 соответственно (рис. 10). Также, последовательность с идентификатором JQ1984, которая с помощью старых алгоритмов ошибочко классифицировалась как H3.Y, настоящей версии является H3.3. Общая статистика представлена в таблице 2.

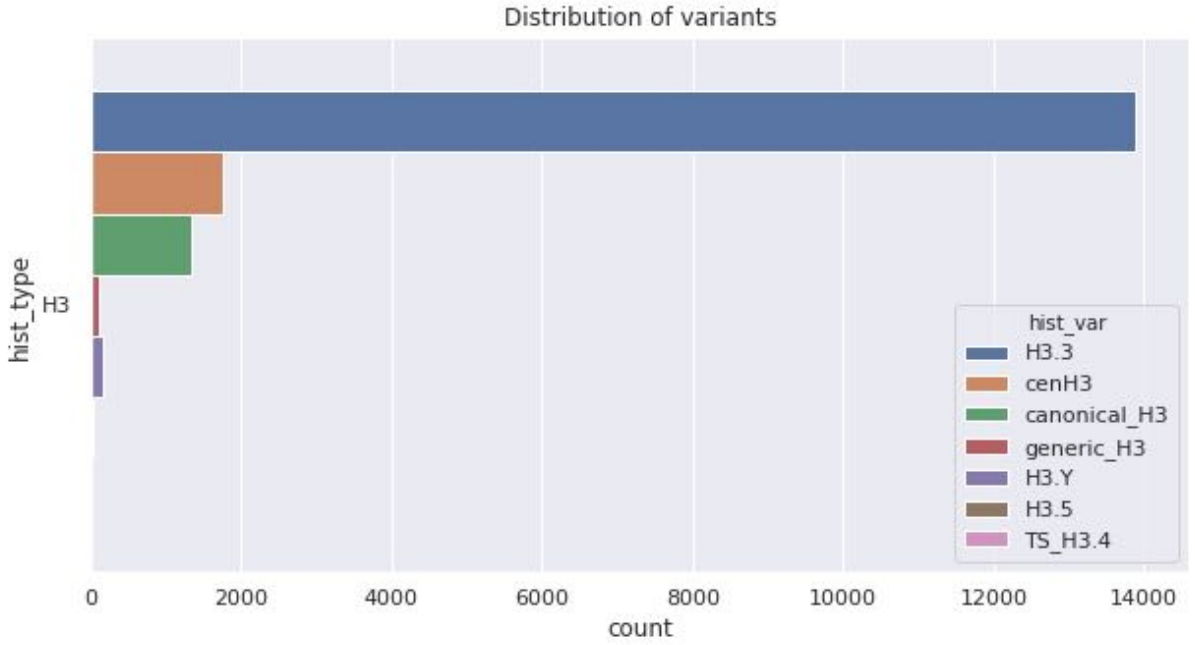

Рисунок 10. Распределение аминокислотных последовательностей гистонов H3 у грибов в базе данных HistoneDB 3.0

Тип гистона	Общее количество	Количество курируемых последовательностей	Количество последовательностей, полученных автоматически
H1	16847	57	16790
H2A	32198	150	32048
H2B	28637	56	28581
H3	84094	70	84024
H4	24334	14	24320

Табл. 2. Количество аминокислотных последовательностей в базе данных

3.3 Клиент-серверное приложение

База данных HistoneDB 3.0 оснащена клиент-серверным приложением (https://histonedb.bioeng.ru/), которое позволяет взаимодействовать с аминокислотными последовательностями гистонов и проводить их анализ.

Основным функциональным элементом является просмотр содержания базы данных. Для каждого типа и подсемейства гистонов можно увидеть таблицу аминокислотных последовательностей. Каждая ее запись включает идентификатор (ACCESSION VERSION), тип и название подсемейства гистоновых белков, вид, тип и класс живого организма, а также значения метрики качества классификации аминокислотной последовательности для автоматически собранных последовательностей. Благодаря наличию внешних ссылок, пользователь имеет возможность перейти по на запись о последовательности в NCBI или запись об организме в NCBI Taxonomy. Сервис позволяет сортировать, фильтровать, выбирать и откладывать последовательности базы данных в корзину для дальнейшего анализа.

Еще одним важным функциональным элементом является возможность строить и визуализировать множественные выравнивания с указанием всех свойств и особенностей семейства, а также просматривать филогенетические деревья и филетическое распределение. Пользователь может просмотреть выравнивания последовательностей курируемого набора, а также выбирать любые последовательности из базы данных для построения отдельного выравнивания. Для каждого типа гистона на главной вкладке иллюстрируется филогенетическое дерево, построенное для последовательностей курируемого набора. А для каждого варианта можно увидеть интерактивную диаграмму филетического распределения аминокислотных последовательностей.

Более того, сервис позволяет классифицировать собственную последовательность. В результате пользователь получает не только самый вероятное подсемейство гистонов, к которому принадлежит исследуемая

последовательность, но и сравнить ее с имеющимися в базе данных другими последовательностями.

Основным нововведением в клиент-серверное приложение стало добавление раздела с таблицей, которая содержит все гены гистонов человека. Она включает название и идентификаторы гена (HGNC symbol, NCBI gene ID, Ensemble gene ID), идентификатор транскрипта (Transcript stable ID), идентификатор мРНК (RefSeq mRNA ID), идентификатор белка (RefSeq peptide ID), а также длину белка в аминокислотах, названия типа и подсемейства гистонов, название изоформы и ссылки на источники.

Глава 4. Анализ разнообразия и классификация последовательностей белков гистонов

4.1 Систематизация разнообразия гистоновых белков в широком спектре живых организмов

В силу своего многообразия гистоновые белки делятся на различные семейства. Однако, до сих пор отсутствует систематизированное представление о значимых вариациях аминокислотных последовательностей гистонов и их структурных особенностях. Для того, чтобы более детально изучить каждое семейство гистоновых вариантов, был сформирован новый курируемый набор гистоновых белков, на основе которого проведен биоинформатический анализ их аминокислотных последовательностей, а также была разработана система классификации гистонов на подсемейства, которая учитывает внутривидовые различия.

Основываясь на литературных данных, можно утверждать, что различные подсемейства гистонов имеют широкое функциональное разнообразие. Однако, существующие на сегодняшний день подходы к классификации гистонов выделяют лишь порядка 35 различных подсемейств, в число которых включаются как отдельные варианты, специфичные некоторым видам живых организмов, так варианты, которые представлены в в целом ряде видов живых организмов. Например, варианты H2A.W и OO H1.8 присутствуют исключительно в растениях и в ооцитах млекопитающих, соответственно [7,35,37,45–47]. А вариант H2A.X присутствует как у многоклеточных, так и у растений, грибов и протистов. При этом известно, что у большинства грибов вариант H2A.X выполняет функцию канонических гистонов, которые у него отсутствуют [24]. Кроме того, функциональные различия гистоновых белков могут зависеть от сплайс-изоформы. Например, у человеческий ген macroH2A.1 кодирует 2 сплайс-изоформы, одна из которых может связывать производные метаболиты NAD+, такие как АДФ-рибоза и поли-АДФ-рибоза, а другая нет [36,48].

Отсюда видно, что существуют белки одного подсемейства гистонов из разных видов живых организмов, которые имеют множество отличий на первичном и структурном уровнях, а также могут выполнять различные или схожие функции. Кроме того, отдельного внимания заслуживают канонические гистоновые белки. Несмотря на свою высокую консервативность, в них выделяют отдельные изоформы, которые могут играть разную функциональную роль.

Для того, чтобы расширить наше представление о многообразии гистоновых белков, нами была разработана новая система классификации гистонов. В ее основу легла идея иерархического подхода. Для каждого известного подсемейства гистонов мы сгруппировали информацию об изученных изоформах, которые отличаются не только на уровне разных видов, но и внутри них. В качестве примера, на рисунке 11 проиллюстрирована схема иерархической классификации гистонов семейства H2A. Благодаря такому представлению, мы можем отметить, что разнообразие канонических изоформ не менее широко, чем разнообразие вариантов подсемейства.

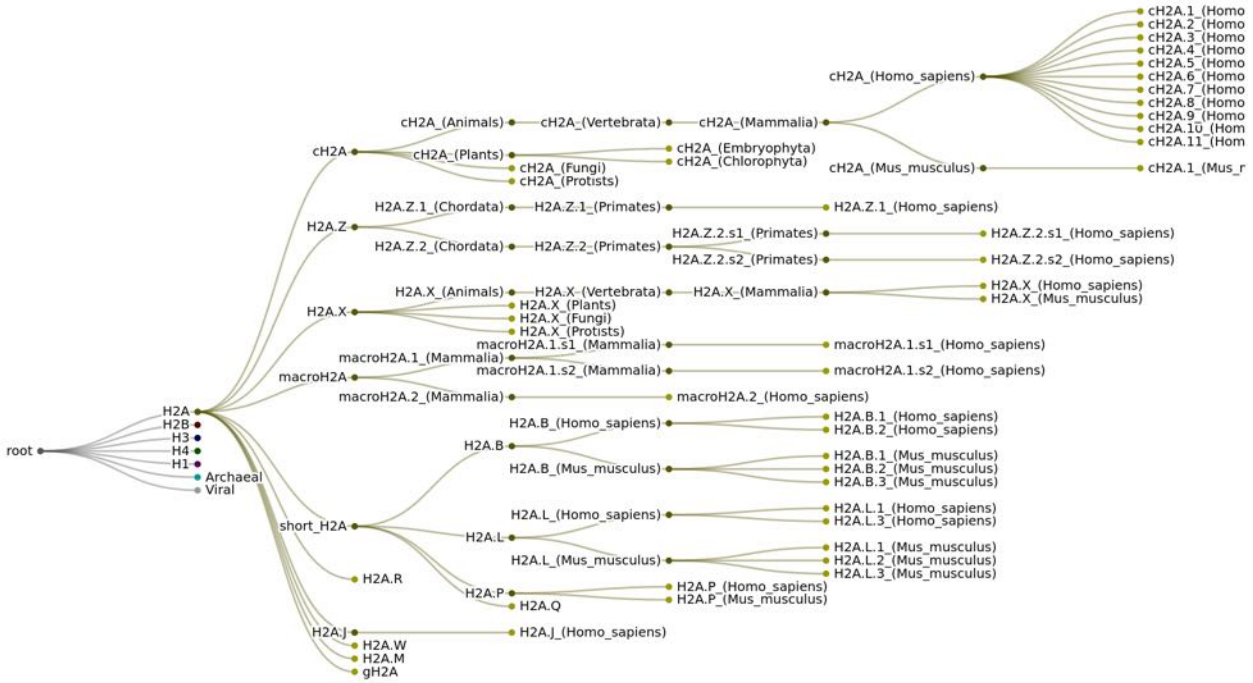

Рисунок 11. Иерархическая система классификации гистоновых белков семейства H2A.

4.2 Кластеризация последовательностей гистоновых белков

Для каждого семейства гистоновых белков мы провели кластеризацию аминокислотных последовательностей, в ходе которой были получены филогенетическое дерево и матрица попарной идентичности между последовательностями. Анализ результатов продемонстрировал, что семейства гистонов H2A отчетливо выделяются два крупных подсемейства (рис. 11). Один из них включает в себя группу “коротких” гистонов H2A (short H2A), состоящую из подгрупп гистонов H2A.P, H2A.B, H2A.L и H2A.Q [33,49,50]. Второе подсемейство объединяет последовательности канонических белков и других вариантов H2A: H2A.X, H2A.Z, macroH2A, H2A.W и H2A.R [1,7,29,33,35,48]. Опираясь на матрицу попарной идентичности (рис. 12) мы видим, что полученные подсемейства отличаются степенью консервативности аминокислотных последовательностей. Наиболее высокую консервативность демонстрируют канонические гистоны семейства H2A (более 80% идентичности). При этом группу “коротких” H2A составляют последовательности с очень низкими показателями консервативности по сравнению с остальными вариантами семейства H2A (менее 46% идентичности во всей группе). Наименее консервативным является H2A.P. Интересно отметить, что несмотря на низкую степень идентичности, в некоторых подгруппах “коротких” H2A гистонов можно заметить отдельные кластеры последовательностей (рис. 12) .

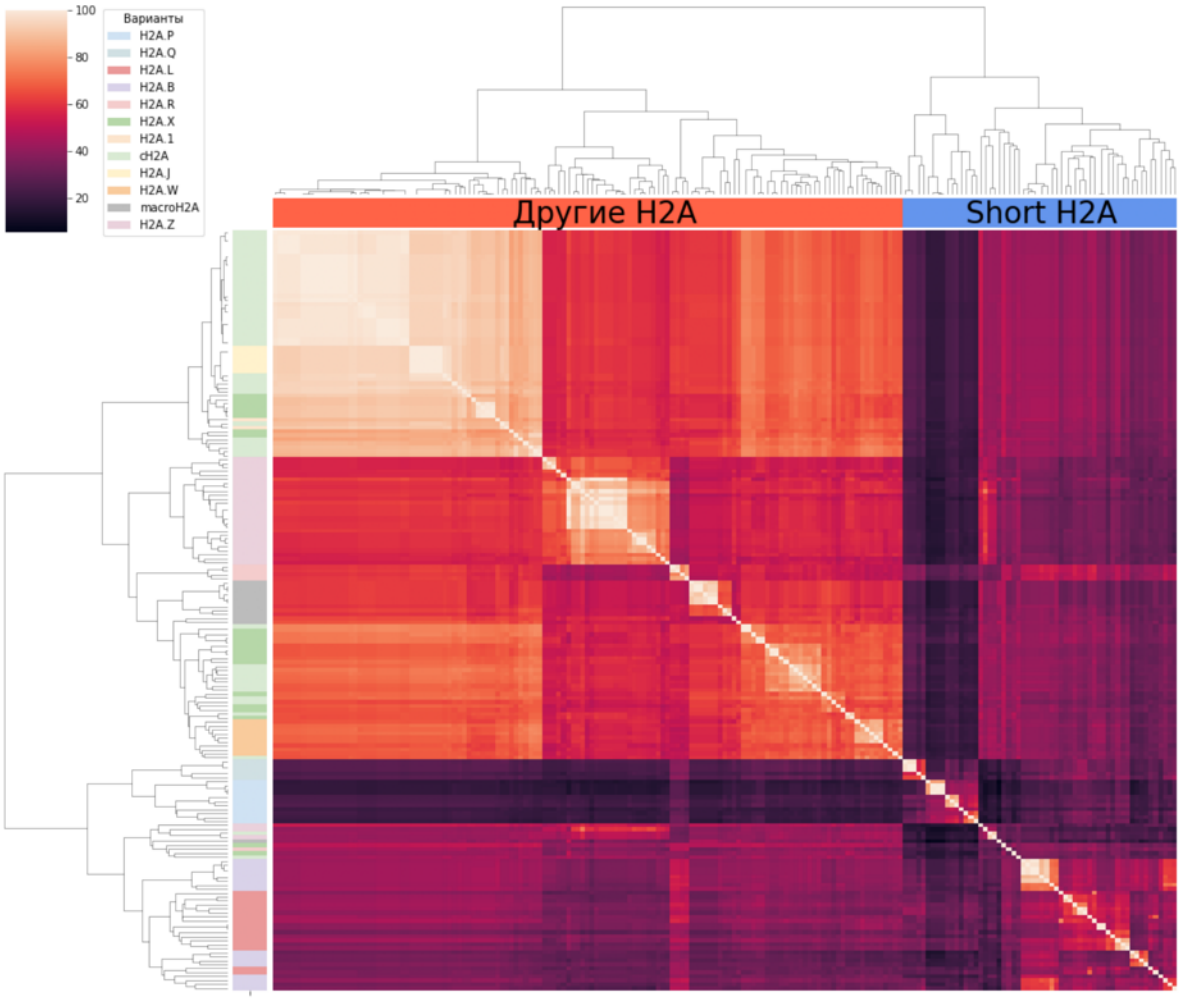

Рисунок 12. Матрица попарной идентичности между аминокислотными последовательностями гистонов семейства H2A, полученная в результате кластеризации гистоновых белков H2A. Цветовая шкала отражает степень идентичности аминокислотных последовательностей. Чем более темный оттенок, тем ближе значение к нулю и ниже идентичность. Дерево иерархической кластеризации гистоновых белков H2A представлено слева и сверху матрицы. Цветовая шкала слева определяет подсемейство, к которому относится последовательность. Цветовая шкала сверху определяет два крупных кластера гистоновых белков семейства H2A: “короткие” H2A и другие варианты семейства H2A.

4.3 Филогенетический анализ и классификация гистоновых белков

Для проведения филогенетического анализа мы использовали алгоритмы максимального правдоподобия для построения филогенетического дерева. Так как C- и N-хвосты гистоновых белков являются неупорядоченными, мы использовали только аминокислотные последовательности глобулярных доменов. Построенное филогенетическое дерево (рис. 13а) позволяет отметить, что самыми дивергентными гистонами семейства являются варианты “коротких” H2A, а самыми близкими к ним ортологами являются гистоны недавно обнаруженного варианта H2A.R [46]. Опираясь на структуры дерева, интересно отметить, что белки канонического H2A и варианта H2A.X скорее всего развивались отдельно друг от друга в процессе эволюции несколько раз. Кроме того, так как вариант H2A.X играет важную роль в репарации повреждений ДНК вероятно, вероятно он предшествовал канонической форме H2A [1]. Более того, мы можем заметить на дереве отдельные клады внутри вариантов H2A.B, H2A.P, H2A.Q, H2A.L, которые принадлежат подсемейству “коротких” гистонов семейства H2A (рис. 13б). Для того, чтобы охарактеризовать обнаруженные подгруппы, а также сопоставить результаты филогенетического анализа с результатами кластеризации последовательностей, мы отобрали только аминокислотные последовательности “коротких” H2A гистонов и кластеризовали их с помощью иерахического метода, как и на первом этапе исследования. Анализ матрица попарной идентичности между аминокислотными последовательностями “коротких” H2A (рис. 12в) демонстрирует наличие нескольких кластеров внутри аминокислотных последовательностей вариантов H2A.B и H2A.Q. Последовательности в этих кластерах характеризуются достаточно высокой консервативностью: в двух кластерах H2A.B идентичность поставила 59% и 74%, а в трех кластерах H2A.Q - 66%, 86% и 95%. Сравнив полученные результаты с филогенетическим анализом нам удалось установить, что каждый кластер представляет собой набор аминокислотных последовательностей, которые принадлежат к одной или

нескольким кладам филогенетического дерева. Иными словами, результаты филогенетического анализа аминокислотных последовательностей гистоновых белков без неупорядоченных C- и N-хвостов согласуются с результатами кластеризации аминокислотных последовательностей, включающих гистоновые хвосты (рис. 13б).

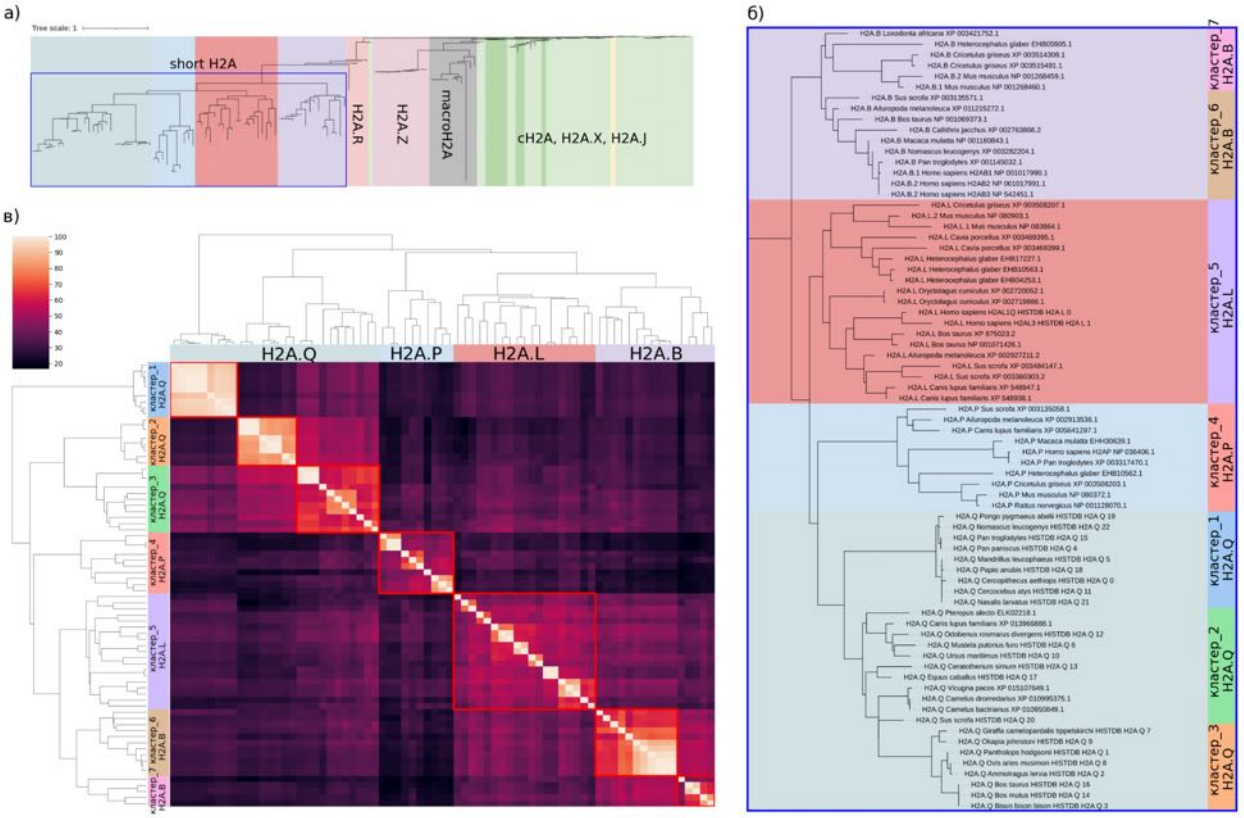

Рисунок 13. Филогенетический анализ аминокислотных последовательностей гистонов семейства H2A. а) Филогенетическое дерево, построенное с помощью методов максимального правдоподобия, для аминокислотных последовательностей глобулярных доменов (HFD) гистонов H2A. б) Визуализация крупным планом ветвей филогенетического дерева с последовательностями вариантов H2A.B, H2A.P, H2A.Q, H2A.L. Справа на дереве отмечены названия кластеров, полученные в результате кластеризации исходных последовательностей “коротких” гистонов семейства H2A, включающих неупорядоченные хвосты. в) Матрица попарной идентичности между аминокислотными последовательностями “коротких” гистонов семейства H2A, полученная в ходе кластеризации “коротких” гистоновых

белков H2A. Цветовая шкала слева определяет кластеры “коротких” гистоновых белков H2A. Цветовая шкала сверху определяет название варианта семейства H2A, к которому относится последовательность.

4.4 Вариации последовательностей гистоновых белков и их влияние на структуру нуклеосомы

Чтобы охарактеризовать особенности каждого из подсемейств гистоновых белков, были проанализированы множественные выравнивания их аминокислотных последовательностей. В результате были выявлены различные вариации, являющиеся структурно и функционально значимыми. Важно отметить, что несмотря на высокую консервативность и одинаковую функциональную роль канонических гистонов, среди их последовательностей также были обнаружены некоторые отличия. Например, аминокислотные последовательности канонического H2A у некоторых протистов (Alveolata) характеризуются длинным N-концевым хвостом, а у растений - немного удлиненным C-концевым хвостом. Интересно, что гистоны грибов сильно отличаются от гистонов остальных видов живых организмов. При этом у них отсутствуют две аминокислоты, находящиеся в области кислотного лоскута (acidic patch), что может приводить к снижению стабильности хроматина [17].

Как показывает анализ множественных выравниваний, вариант H2A.X является наиболее близким к канонической форме гистонов (идентичность аминокислотных последовательностей более 71%). Аминокислотная последовательность гистоновой складки варианта H2A.X отличается от канонического H2A всего в пяти позициях, а область структурного домена “docking domain” - в двух позициях. Вариант H2A.X также интересен тем, что у него на C-концевом хвосте имеется мотив SQE/DФ (SQAY у дрозофилы), где Ф- представляет собой гидрофобный остаток. Данная особенность имеет важную функциональную роль. Известно, что при фосфорилировании остатка серина (S) вариант H2A.X способен служить маркером

двухцепочечных разрывов ДНК и тем самым привлекать машинерию, устраняющую повреждение ДНК [51].

Вариант H2A.Z представляет собой множество последовательностей, обладающих высокой консервативностью между собой (идентичность последовательностей более 72%). При этом анализ выравниваний аминокислотных последовательностей варианта H2A.Z у хордовых позволил выделить две группы белков, которые отличаются всего тремя аминокислотными остатками. Несмотря на столь незначительные вариации, некоторые исследования свидетельствуют о том, что данные группы могут быть функционально независимыми. Например, белки варианта H2A.Z, попавшие в одну из групп, лучше взаимодействует с белком, содержащим бромодомен BRD2 [28]. Для белков другой группы было обнаружено, что они предпочтительно связывается с гистоном семейства H3, триметилированным по лизину 4 (H3K4me3) [29].

Самым длинным гистоновым вариантом является macroH2A (в среднем 360 аминокислотных остатков). Такая длина объясняется наличием негистонового макродомена, который соединен с С-концом центральной области глобулярного домена. Аминокислотные последовательности варианта macroH2A демонстрируют высокий уровень консервативности между собой (более 64% идентичности по всей длине). При этом область негистонового макродомена очень разнообразна, что может свидетельствовать о его структурной и функциональной значимости. Более того, на человеческих белках macroH2A было показано, что некоторые из них способны связывать производные метаболиты NAD+, такие как АДФ-рибоза и поли-АДФ-рибоза, с помощью макродоменов [48].

Аминокислотные последовательности варианта H2A.W, обнаруженного исключительно в растениях, являются представителями одного из самых консервативных вариантов семейства H2A (более 76% идентичности). Наиболее разнообразные вариации встречаются в областях, расположенных между альфа-спиралями. Вариант H2A.W также может быть охарактеризован

наличием специфичного мотива SPKK, расположенного в С-концевом хвосте. Исходя из литературных данных, можно предположить, что данный мотив играет важную роль в структуре хроматина, так как он может связываться с малой бороздкой ДНК [34].

Отдельного внимания заслуживают “короткие” гистоны семейства H2A, которые преимущественно экспрессируются во время развития мужских половых клеток млекопитающих [49]. Недавно было обнаружено, что они обладают рядом особенностей, схожими с мутациями онкогистонов, которые могут приводить к дестабилизации нуклеосом [49]. Кроме того, что “короткие” H2A обладают самой низкой консервативностью (рис. 14), у них обнаружены значимые вариации в аминокислотных последовательностях. Например, все варианты “коротких” H2A имеют укороченные С-концевой хвост и структурный домен “docking domain”, характерный для всех белков семейства H2A. При этом H2A.Q имеет наиболее короткий хвост (рис. 14). Важно отметить, что в аминокислотных последовательностях гистонов группы “коротких” H2A имеются вариации в сайтах связывания с ДНК и в регионе “acidic patch”, который отвечает за стабильность хроматина [17]. Например, у варианта H2A.P отсутствуют оба аргина, участвующие в связи с малой бороздкой ДНК, а также четыре из шести остатков кислотного лоскута, причем один из них заменен на положительно заряженную аминокислоту - аргинин. Вариант H2A.L лишен всего лишь трех остатков из региона “acidic patch”. В аминокислотных последовательностях H2A.B обнаружены всего два из шести остатков кислотного лоскута, и один из аргининов в стайте связывания с ДНК. Интересно отметить, что наблюдаются отличия между двумя кластерами последовательностей в группе гистоновых вариантов H2A.B, которые были получены в ходе кластерного анализа “коротких” гистонов семейства H2A. Например, можно выделить две позиции, находящиеся в регионе “acidic patch”. В одной из них оказалась положительно заряженная аминокислота в обоих кластерах H2A.B (E127K и E127R), в другой - только в одном из кластеров (E159R). Аминокислотные

последовательности недавно обнаруженного H2A.Q также имеют различия между выявленными кластерами [45]. В одном из них отсутствуют все кислотные остатки в регионе “acidic patch” (причем в трех позициях встречаются положительно заряженные аминокислоты: E122K, E130K, E158K) и все аргинины в сайтах связывания с ДНК. В двух других кластерах последовательностей у варианта H2A.Q можно увидеть только один из шести остатков кислотного лоскута.

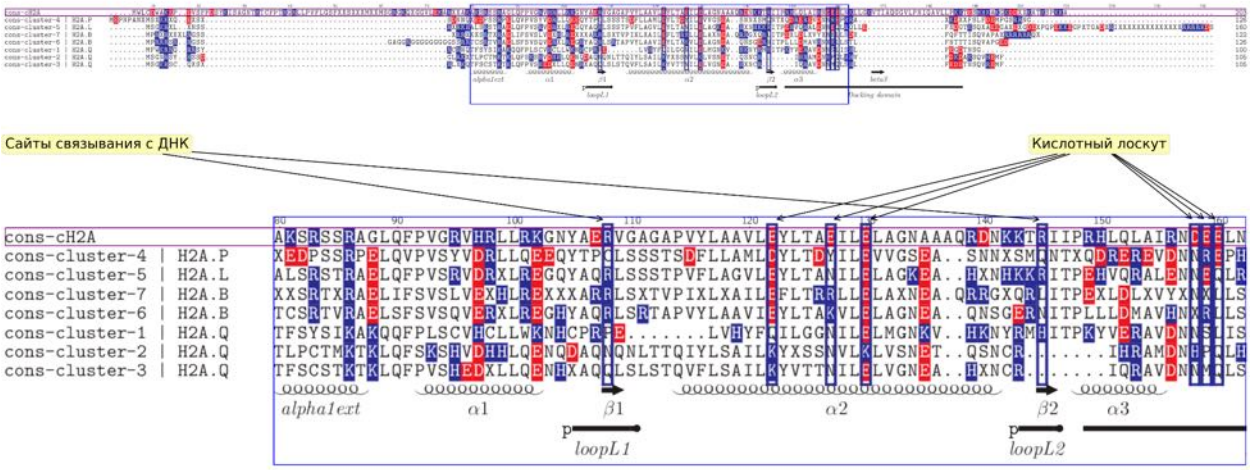

Рисунок 14. Множественное выравнивание аминокислотных последовательностей “коротких” гистонов семейства H2A. Выравнивание построено для консенсусных последовательностей, полученных путем множественного выравнивания для каждого отдельного кластера, полученного в ходе кластерного анализа “коротких” гистоновых вариантов H2A. Синей прямоугольной рамкой выделены аминокислоты кислотного лоскута и сайты связывания с ДНК (характерной чертой всех гистонов семейства H2A является наличие аргинина в сайтах связывания с ДНК). Синим цветом закрашены аминокислотные остатки, заряженные положительно. Красным цветом закрашены аминокислотные остатки, заряженные отрицательно.

4.5 Обсуждение результатов

Комплексный анализ литературы и биоинформатических баз данных позволил нам собрать новый курируемый набор аминокислотных последовательностей гистоновых белков. В него вошли последовательности из 154 живых организмов, включая археи и вирусы. Ввиду широкого видового разнообразия гистонов, мы также приняли решение разработать новую систему классификации гистонов. Ее ключевой особенностью является использование иерархического представления семейств и подсемейств гистоновых белков, в результате которого мы можем выявить различные функционально значимые особенности гистонов не только на уровне видов, но и внутри них.

В результате биоинформатического анализа разнообразия гистоновых белков нового курируемого набора, мы охарактеризовали вариации аминокислотных последовательностей, а также структурные и функциональные особенности различных подсемейств гистоновых белков. В результате мы показали, что каждое отдельное подсемейство обладает специфичным набором характеристик, которые могут располагаться в потенциально функционально значимых сайтах последовательности и влиять на физико-химические свойства нуклеосомы. При этом важную роль в динамике хроматина могут играть не только масштабные отличия (например, наличие специфичных доменов или мотивов), но и отдельные точечные вариации в аминокислотных последовательностях (например, вариации, расположенные в кислотном лоскуте или в сайтах взаимодействия с другими гистонами и ДНК). Например, в результате проведенного в данной работе биоинформатического анализа гистоновых белков нового курируемого набора были построены филогенетические деревья, которые позволили обнаружить новые ранее неизвестные подсемейства коротких гистонов H2A. Мы также охарактеризовали их отличия, в числе которых присутствуют потенциально функционально значимые вариации, расположенные в области кислотного лоскута и в сайтах связывания с ДНК. На основании полученных результатов

мы можем предположить, что выявленные подсемейства коротких гистонов H2A могут играть важную функциональную роль в изменении стабильности нуклеосомы.

Новый курируемый набор аминокислотных последовательностей гистоновых белков был использован также для расширения и модернизации базы данных HistoneDB. Не нем были обучены скрытые марковские модели, лежащие в основе программы HMMER, а также из них была собрана база данных BLAST, которая используется программой BLASTP. Благодаря разделению алгоритма поиска и классификации гистоновых белков на два ключевых этапа, мы смогли достичь более точной классификации гистонов на подсемейства. Оба этапа реализуются с применением программ для поиска гомологичных последовательностей HMMER и BLASTP, которые направлены на определение степени идентичности последовательностей сравниваемых белков. В силу того, что программа HMMER основана на использовании статистических моделей и скрытых марковских процессах, она хорошо справляется с обнаружением дальних гомологов, что позволяет отличить семейства гистонов. Однако, ее недостаток в том, что она не справляется со строгой классификацией на гистоновые подсемейства. В то же время, существует программа BLASTP, которая оценивает сходство последовательностей, опираясь на матрицу аминокислотных замен BLOSUM, что позволяет уделить больше внимания более специфичным отличиям между аминокислотными последовательностями различных подсемейств.

Кроме того, на первом этапе обучение скрытых марковских моделей проводилось с использованием не всей аминокислотной последовательности белка, а только с последовательностью его глобулярного домена. Благодаря такому подходу нам удалось исключить влияние аминокислотной последовательности хвостов на результат классификации по типам гистонов. Важно отметить также, что алгоритмы классификации стали более гибкими за счет проведения уточнения классификации в конце второго этапа путем рассмотрения мотивов и доменов, специфичных для отдельных подсемейств.

Таким образом, в ходе разработки обновленной версии базы данных гистоновых белков HistoneDB 3.0 мы не только увеличили количество хранящихся в ней последовательностей в 2,5 раза (более 186 тысяч последовательностей гистонов), но и усовершенствовали алгоритмы поиска и классификации гистоновых белков на различные подсемейства. Вместе с тем мы разработали новый веб-интерфейс для взаимодействия с базой данных, который обеспечил стабильность функциональных элементов и позволил провести более комплексный анализ гистоновых белков в широком спектре живых организмов, в том числе в рахеях и вирусах. Благодаря модернизированной базе данных HistoneDB 3.0, в данной диссертационной работе продемонстрировано, что несмотря на то, что гистоновые белки считаются эволюционно высоко консервативными, некоторые подсемейства содержат белки, у которых идентичность аминокислотных последовательностей в области глобулярных доменов ниже 46%.

Заключение

Ключевую роль в изучении процессов, лежащих в основе функционирования генома и жизнедеятельности клетки, играет динамика хроматина, а также гистоновые белки, плотно ассоциированные с ДНК. В зависимости от различных режимов нуклеосомы, влияющих на ее компактность, и факторов, влияющих на структурную динамику нуклеосомы, меняется тонкая взаимная регуляция активности генов, которая позволяет организму функционировать и развиваться.

В данной работе проведена систематизация знаний о гистоновых белках и их роли в структуре хроматина. Для более комплексного понимания многообразия гистонов, а также выявления различных функционально значимых особенностей аминокислотных последовательностей их подсемейств, разработана новая система классификации гистоновых белков, которая учитывает, не только видовое, но и внутривидовое разнообразие гистонов. Для расширения и улучшения стабильности биоинформатических баз данных, использованных в данном исследовании, проведена модернизация функциональных инструментов и разработаны новые более точные и гибкие алгоритмы поиска и классификации гистоновых белков в широком спектре живых организмов, включая археи и вирусы.

Результаты данной исследовательской работы планируются к публикации в рецензируемых журналах, в том числе международного уровня с высоким импакт-фактором. В данный момент результаты биоинфоратического анализа разнообразия гистоновых белков семейства H2A отправлены на публикацию в Вестник Московского университета (Серия 16. Биология.). Кроме того, планируются публикации, в которых будут описаны результаты разработки нового подхода к классификации гистоновых белков на семейства и подсемейства, учитывающий видовое и межвидовое разнообразие, а также результаты расширения базы данных HistoneDB и разработки новых более точных и гибких алгоритмов автоматического поиска и классификации гистоновых белков.

Выводы диссертационной работы:

1. Собран новый курируемый набор аминокислотных последовательностей гистоновых белков из 154 различных живых организмов, который включает более, чем 50 различных подсемейств гистонов.
2. Разработана новая система классификации гистоновых белков, учитывающая как видовое, так и внутривидовое разнообразие гистонов, которая позволила выявить новые функционально значимые подсемейства и охарактеризовать их структурные и физико-химические свойства, а также легла в основу расширения базы данных HistoneDB и улучшения алгоритмов поиска и классификации гистоновых белков в эукариотах, археях и вирусах.
3. В результате биоинформатического анализа аминокислотных последовательностей гистоновых белков обнаружены ранее неизвестные подсемейства коротких вариантов H2A, в которых идентичность последовательностей составила более 52%, а также обнаружены отличия в сайтах, расположенных в области кислотного лоскута, что может влиять на стабильность нуклеосомы.
4. В ходе модернизации базы данных гистоновых белков, разработана новая расширенная версия HistoneDB 3.0, которая содержит более 186 тысяч последовательностей, классифицированных на более, чем 50 различных подсемейств гистонов, а также позволяет проводить эволюционный анализ гистоновых белков в широком спектре живых организмов.

Благодарности

Автор выражает благодарность

научному руководителю д.ф.-м.н. Шайтану Алексею Константиновичу;

Грибкову Анну Кирилловну и Неугодова Артема Михайловича за помощь в решении некоторых задач для обновления базы данных HistoneDB.

Работа выполнена при поддержке гранта Президента РФ МД-1131.2022.1.4.

Список работ, опубликованных автором по теме диссертации

1. Singh-Palchevskaiia L., Shaytan A. K. Development of algorithms for mining and classification of histone proteins //Biophysical Journal. – 2022. – Т. 121. – №. 3. – С. 361a.
2. Сингх-Пальчевская Л., Шайтан А.К. Кластеризация И Анализ Последовательностей Гистоновых Белков H2a. Новосибирский национальный исследовательский государственный университет, 2022.
3. Сингх-Пальчевская Л., Шайтан А.К. Филогенетический анализ и классификация коротких H2A-гистонов. Кубанский государственный технологический университет, 2023.
4. Сингх-Пальчевская Л., Шайтан А. К. Разнообразие гистонов H2A и их влияние на структурные свойства нуклеосомы // Вестник Московского университета. Серия 16. Биология. 2023. (отправлена на публикацию).

Литература

1. Talbert P.B., Henikoff S. Histone variants at a glance // J. Cell Sci. 2021. Vol. 134, № 6. P. jcs244749.
2. Draizen E.J. et al. HistoneDB 2.0: a histone database with variants—an integrated resource to explore histones and their variants // Database. 2016. Vol. 2016. P. baw014.
3. Luger K. et al. Crystal structure of the nucleosome core particle at 2.8 A resolution // Nature. 1997/09/26 ed. Macmillan Magazines Ltd., 1997. Vol. 389, № 6648. P. 251–260.
4. Kouzarides T. Chromatin Modifications and Their Function // Cell. Elsevier, 2007. Vol. 128, № 4. P. 693–705.
5. Malik H.S., Henikoff S. Phylogenomics of the nucleosome // Nat. Struct. Biol. 2003. Vol. 10, № 11. P. 882–891.
6. Talbert P.B. et al. A unified phylogeny-based nomenclature for histone variants // Epigenetics Chromatin. 2012. Vol. 5, № 1. P. 7.
7. Kawashima T. et al. Diversification of histone H2A variants during plant evolution // Trends Plant Sci. Elsevier, 2015. Vol. 20, № 7. P. 419–425.
8. Henneman B. et al. Structure and function of archaeal histones // PLOS Genet. Public Library of Science, 2018. Vol. 14, № 9. P. e1007582.
9. Talbert P.B., Armache K.-J., Henikoff S. Viral histones: pickpocket’s prize or primordial progenitor? // Epigenetics Chromatin. 2022. Vol. 15, № 1. P. 21.
10. Seal R.L. et al. A standardized nomenclature for mammalian histone genes // Epigenetics Chromatin. 2022. Vol. 15, № 1. P. 34.
11. Marzluff W.F. Metazoan replication-dependent histone mRNAs: a distinct set of RNA polymerase II transcripts // Curr. Opin. Cell Biol. 2005. Vol. 17, № 3. P. 274–280.
12. Marzluff W.F., Wagner E.J., Duronio R.J. Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail // Nat. Rev. Genet. 2008. Vol. 9, № 11. P. 843–854.
13. Lyons S.M. et al. A subset of replication-dependent histone mRNAs are

expressed as polyadenylated RNAs in terminally differentiated tissues // Nucleic Acids Res. 2016. P. gkw620.

14. Marzluff W.F., Koreski K.P. Birth and Death of Histone mRNAs // Trends Genet. 2017. Vol. 33, № 10. P. 745–759.
15. Eirín-López J.M. et al. Long-Term Evolution of Histone Families: Old Notions and New Insights into Their Mechanisms of Diversification Across Eukaryotes // Evolutionary Biology / ed. Pontarotti P. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. P. 139–162.
16. Marzluff W.F. et al. The Human and Mouse Replication-Dependent Histone Genes // Genomics. 2002. Vol. 80, № 5. P. 487–498.
17. Shaytan A.K., Landsman D., Panchenko A.R. Nucleosome adaptability conferred by sequence and structural variations in histone H2A-H2B dimers // Curr. Opin. Struct. Biol. 2015/03/04 ed. 2015. Vol. 32. P. 48–57.
18. Armeev G.A. et al. Linking chromatin composition and structural dynamics at the nucleosome level // Curr. Opin. Struct. Biol. 2019. Vol. 56. P. 46–55.
19. Franklin S.G., Zweidler A. Non-allelic variants of histones 2a, 2b and 3 in mammals: 5599 // Nature. Nature Publishing Group, 1977. Vol. 266, № 5599. P. 273–275.
20. Huynh L.M., Shinagawa T., Ishii S. Two Histone Variants TH2A and TH2B Enhance Human Induced Pluripotent Stem Cell Generation // Stem Cells Dev. 2016. Vol. 25, № 3. P. 251–258.
21. Zalensky A.O. et al. Human Testis/Sperm-specific Histone H2B (hTSH2B): MOLECULAR CLONING AND CHARACTERIZATION // J. Biol. Chem. Elsevier, 2002. Vol. 277, № 45. P. 43474–43480.
22. Shinagawa T. et al. Histone Variants Enriched in Oocytes Enhance Reprogramming to Induced Pluripotent Stem Cells // Cell Stem Cell. 2014. Vol. 14, № 2. P. 217–227.
23. Zambrano-Mila M.S., Aldaz-Villao M.J., Armando Casas-Mollano J. Canonical Histones and Their Variants in Plants: Evolution and Functions // Epigenetics in Plants of Agronomic Importance: Fundamentals and Applications

/ ed. Alvarez-Venegas R., De-la-Peña C., Casas-Mollano J.A. Cham: Springer International Publishing, 2019. P. 185–222.

24. Freitag M. 5 Fungal Chromatin and Its Role in Regulation of Gene Expression // Fungal Genomics / ed. Nowrousian M. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014. P. 99–120.
25. Dalmasso M.C. Canonical and variant histones of protozoan parasites // Front. Biosci. 2011. Vol. 16, № 1. P. 2086.
26. Константинович Ш.А. Интегративное моделирование структуры и динамики биомакромолекулярных комплексов: доктор наук. ФГБОУ ВО «Московский государственный университет имени М.В. Ломоносова», 2021. P. 488.
27. Talbert P.B., Henikoff S. Histone variants — ancient wrap artists of the epigenome: 4 // Nat. Rev. Mol. Cell Biol. Nature Publishing Group, 2010. Vol. 11, № 4. P. 264–275.
28. Draker R. et al. A Combination of H2A.Z and H4 Acetylation Recruits Brd2 to Chromatin during Transcriptional Activation // PLOS Genet. Public Library of Science, 2012. Vol. 8, № 11. P. e1003047.
29. Dryhurst D. et al. Characterization of the histone H2A.Z-1 and H2A.Z-2 isoforms in vertebrates // BMC Biol. 2009. Vol. 7, № 1. P. 86.
30. Giaimo B.D. et al. The histone variant H2A.Z in gene regulation // Epigenetics Chromatin. 2019. Vol. 12, № 1. P. 37.
31. Adam M. et al. H2A.Z is required for global chromatin integrity and for recruitment of RNA polymerase II under specific conditions // Mol. Cell. Biol. 2001. Vol. 21, № 18. P. 6270–6279.
32. Millar C.B. Organizing the genome with H2A histone variants // Biochem. J. 2013. Vol. 449, № 3. P. 567–579.
33. Molaro A., Young J.M., Malik H.S. Evolutionary origins and diversification of testis-specific short histone H2A variants in mammals // Genome Res. Cold Spring Harbor Lab, 2018. Vol. 28, № 4. P. 460–473.
34. Churchill M.E., Suzuki M. “SPKK” motifs prefer to bind to DNA at

A/T-rich sites // EMBO J. 1989/12/20 ed. 1989. Vol. 8, № 13. P. 4189–4195.

35. Yelagandula R. et al. The Histone Variant H2A.W Defines Heterochromatin and Promotes Chromatin Condensation in Arabidopsis // Cell. Elsevier, 2014. Vol. 158, № 1. P. 98–109.

36. Sun Z., Bernstein E. Histone variant macroH2A: from chromatin deposition to molecular function // Essays Biochem. 2019. Vol. 63, № 1. P. 59–74.

37. Ueda K. et al. Male Gametic Cell-specific Histone gH2A Gene of Lilium longiflorum: Genomic Structure and Promoter Activity in the Generative Cell // Plant Mol. Biol. 2005. Vol. 59, № 2. P. 229–238.

38. Baxevanis A.D., Landsman D. Histone Sequence Database: A Compilation of Highly-Conserved Nucleoprotein Sequences // Nucleic Acids Res. 1996. Vol. 24, № 1. P. 245–247.

39. Sullivan S.A. et al. The Histone Database: a comprehensive WWW resource for histones and histone fold-containing proteins // Nucleic Acids Res. 2000. Vol. 28, № 1. P. 320–322.

40. Sullivan S. et al. The Histone Database // Nucleic Acids Res. 2002. Vol. 30, № 1. P. 341–342.

41. Mariño-Ramírez L. et al. The Histone Database: A Comprehensive Resource for Histones and Histone Fold-Containing Proteins // Proteins. 2006. Vol. 62, № 4. P. 838–842.

42. Mariño-Ramírez L. et al. The Histone Database: an integrated resource for histones and histone fold-containing proteins // Database. 2011. Vol. 2011. P. bar048.

43. Draizen E.J. et al. HistoneDB 2.0: a histone database with variants--an integrated resource to explore histones and their variants // Database J. Biol. Databases Curation. 2016. Vol. 2016. P. baw014.

44. Singh-Palchevskaiia L., Shaytan A.K. Development of algorithms for mining and classification of histone proteins // Biophys. J. Elsevier, 2022. Vol. 121, № 3. P. 361a.

45. Tanaka M. et al. A mammalian oocyte-specific linker histone gene H1oo:

homology with the genes for the oocyte-specific cleavage stage histone (cs-H1) of sea urchin and the B4/H1M histone of the frog // Dev. Camb. Engl. 2001. Vol. 128, № 5. P. 655–664.

46. Jiang D. et al. The evolution and functional divergence of the histone H2B family in plants // PLOS Genet. Public Library of Science, 2020. Vol. 16, № 7. P. e1008964.
47. Strickland M. et al. The complete amino-acid sequence of histone H2B(3) from sperm of the sea urchin Parechinus angulosus // Eur. J. Biochem. 1978. Vol. 89, № 2. P. 443–452.
48. Kustatscher G. et al. Splicing regulates NAD metabolite binding to histone macroH2A // Nat. Struct. Mol. Biol. 2005. Vol. 12, № 7. P. 624–625.
49. Jiang X., Soboleva T.A., Tremethick D.J. Short Histone H2A Variants: Small in Stature but not in Function // Cells. 2020. Vol. 9, № 4.
50. Chew G.-L. et al. Short H2A histone variants are expressed in cancer // Nat. Commun. 2021. Vol. 12, № 1. P. 490.
51. Millar C.B. Organizing the genome with H2A histone variants // Biochem. J. 2013. Vol. 449, № 3. P. 567–579.

Приложения

Приложение А - Листинг кода, реализующего алгоритм автоматической классификации гистоновых белков

```
from django.core.management.base import BaseCommand, CommandError
from browse.models import Histone, Variant, Sequence, Score, ScoreHmm,
    ScoreForHistoneType, Feature
from tools.browse_service import *
from tools.test_model import test_model
from Bio import SearchIO, SeqIO, AlignIO #, pairwise2
from Bio.Blast.Applications import NcbiblastpCommandline
from Bio.Blast import NCBIXML
from Bio.Emboss.Applications import NeedleCommandline
import os, sys, io
import re
import logging
from tqdm import tqdm
from datetime import date, datetime
from cProfile import Profile
# This command is the main one in classifying histone sequences using one of the
    two ways^
# by using HMMs constructed based on these alignment to classify the bigger
    database or
# by using BLASTP alignments to classify the bigger database
# see handle() for the workflow description.
class Command(BaseCommand):
    help = 'Build HistoneDB by first loading the seed sequences and then parsing the
    database file'
    # Logging info
    logging.basicConfig(filename=os.path.join(LOG_DIRECTORY,
    "classifyvariants.log"),
```

```
            format='%(asctime)s %(name)s %(levelname)-8s %(message)s',
            level=logging.INFO,
            datefmt='%Y-%m-%d %H:%M:%S')
    log = logging.getLogger(__name__)
    def add_arguments(self, parser):
        parser.add_argument(
        "-f",
        "--force",
        default=False,
        action="store_true",
        help="Force the regeneration of HMM from seeds, HUMMER search in
    db_file, Test models and loading of results to database")
        parser.add_argument(
        "--test",
        default=False,
        action="store_true",
        help="Use this option for test running if you need a small version of
    specified database")
        parser.add_argument(
        "--profile",
        default=False,
        action="store_true",
        help="Profile the command")
    def _handle(self, *args, **options):
self.log.info('==================================================
    ======')
        self.log.info('===          classifyvariants START          ===')
    self.log.info('==================================================
    =======')
        self.start_time = datetime.now()
```

```
        self.log.info('HMMER_PROCS !!!! {}'.format(HMMER_PROCS))
        if 'HMM' in CLASSIFICATION_TYPES:
        if options["force"]:
                ScoreHmm.objects.all().delete()
Sequence.objects.exclude(reviewed=True).update(variant_hmm=None)
                # Determine HMMER thresholds params used to classify sequence
based on HMMER
                self.estimate_hmm_thresholds()
                self.get_scores_for_curated_via_hmm()
        self.classify_via_hmm()
        self.get_stats_hmm()
        if 'BLAST' in CLASSIFICATION_TYPES:
        if options["force"]:
                Score.objects.all().delete()
                Sequence.objects.exclude(reviewed=True).update(variant=None)
                self.get_scores_for_curated_via_blast()
                # self.estimate_thresholds()
                # if options["force"] or not os.path.isfile(self.blast_file + "0"):
                # self.search_blast()
        self.classify_via_blast(force=options["force"])
        self.get_stats_blast()
        seq_num = Sequence.objects.count()
        seqauto_num = Sequence.objects.filter(reviewed=False).count()
        self.log.info(' The database has %d sequences now !!!' % seq_num)
        self.log.info(' %d sequences came from automatic search !!!' %
seqauto_num)
self.log.info('==========================================================
=======')
        self.log.info('===   classifyvariants SUCCESSFULLY finished  ===')
```

```
self.log.info('=========================================================')
def handle(self, *args, **options):
    if options['profile']:
    profiler = Profile()
    profiler.runcall(self._handle, *args, **options)
    profiler.print_stats()
    else:
    self._handle(*args, **options)
# For HMM classification
def estimate_hmm_thresholds(self, specificity=0.95):
    """
    Estimate HMM threshold that we will use for variant classification.
    Construct two sets for every variant:
    negative: The seed alignmnents from every other variant
    positive: the current seed alignment for the variant
    And estimate params from ROC-curves.
    """
    for hist_type_pos, seed_pos in get_seeds():
    variant_name = seed_pos[:-6]
    # Getting all paths right
    positive_seed_aln_file = os.path.join(SEED_DIRECTORY, hist_type_pos,
seed_pos)
    hmm_file = os.path.join(HMM_DIRECTORY, hist_type_pos,
"{}.hmm".format(variant_name))
    output_dir = os.path.join(MODEL_EVALUATION, hist_type_pos)
    if not os.path.exists(output_dir):
            os.makedirs(output_dir)
    positive_examples_file = os.path.join(output_dir,
"{}_postive_examples.fasta".format(variant_name))
```

```
        positive_examples_out = os.path.join(output_dir,
"{}_postive_examples.out".format(variant_name))
        negative_examples_file = os.path.join(output_dir,
"{}_negative_examples.fasta".format(variant_name))
        negative_examples_out = os.path.join(output_dir,
"{}_negative_examples.out".format(variant_name))
        # Unagapping all sequence from seed aln - this will be the positive example
        with open(positive_examples_file, "w") as pf:
                for s in SeqIO.parse(positive_seed_aln_file, "fasta"):
                s.seq = s.seq.ungap("-")
                SeqIO.write(s, pf, "fasta")
        # Searching the positive examples set
        self.search_via_hmm(hmms_db=hmm_file, out=positive_examples_out,
sequences=positive_examples_file, procs=1, E=500)
        # Build negative examples from all other varaints
        with open(negative_examples_file, "w") as nf:
                for hist_type_neg, seed_neg in get_seeds():
                if ((hist_type_pos == hist_type_neg) and (seed_neg == seed_pos)):
                continue
                else:
                sequences = os.path.join(SEED_DIRECTORY, hist_type_neg,
seed_neg)
                for s in SeqIO.parse(sequences, "fasta"):
                s.seq = s.seq.ungap("-")
                SeqIO.write(s, nf, "fasta")
        # Searching through negative example set
        self.search_via_hmm(hmms_db=hmm_file, out=negative_examples_out,
sequences=negative_examples_file, procs=1, E=500)
        # Here we are doing ROC curve analysis and returning parameters
        specificity = 0.8 if ("canonical" in variant_name) else 0.9
```

```
    # Hack to make canoical have a lower threshold and ther variants higher
threshold
    # specificity = 0.05 if ("generic" in variant_name)  else specificity #Hack to
make canoical have a lower threshold and ther variants higher threshold
    parameters = test_model(variant_name, output_dir, positive_examples_out,
negative_examples_out,
                            measure_threshold=specificity)
    # Let's put the parameter data to the database,
    # We can set hist_type directly by ID, which is hist_type_pos in this case -
because it is the primary key in Histone class.
    variant_model = Variant.objects.get(id=variant_name)
    self.log.info("Updating thresholds for {}".format(variant_model.id))
    self.log.info("Threshold = {}, roc_auc = {}".format(parameters["threshold"],
parameters["roc_auc"]))
    variant_model.hmmthreshold = parameters["threshold"]
    variant_model.aucroc = parameters["roc_auc"]
    variant_model.save()
def get_scores_for_curated_via_hmm(self):
    """
    For every curated variant we want to generate a set of scores against HMMs.
    This is needed to supply the same type of information for curated as well as
for automatic seqs.
    """
    # Construct the one big file from all cureated seqs.
    with open(CURATED_GENERICLESS_FASTA, "w") as f,
open(CURATED_GENERIC_FASTA, "w") as fg:
    for hist_type, seed in get_seeds(generic=True):
        seed_aln_file = os.path.join(SEED_DIRECTORY, hist_type, seed)
        for s in SeqIO.parse(seed_aln_file, "fasta"):
        s.seq = s.seq.ungap("-")
```

```
            if 'generic' in seed:
            SeqIO.write(s, fg, "fasta")
            else:
            SeqIO.write(s, f, "fasta")
    # Search all curated except generic by our HMMs
    self.search_via_hmm(hmms_db=COMBINED_HMM_VARIANTS_FILE,
out=CURATED_HISTVAR_RESULTS_FILE,
            sequences=CURATED_GENERICLESS_FASTA, procs=1, E=10)
    ##We need to parse this results file;
    ##we take here a snippet from load_hmmsearch.py, and tune it to work for
our curated seq header format
    for variant_query in
SearchIO.parse(CURATED_HISTVAR_RESULTS_FILE, "hmmer3-text"):
    self.log.info("Loading hmmsearch for variant: {}".format(variant_query.id))
    variant_model = Variant.objects.get(id=variant_query.id)
    for hit in variant_query:
            accession = hit.id.split("|")[1]
            seq = Sequence.objects.get(id=accession)
            try:  # sometimes we get this:    [No individual domains that satisfy
reporting thresholds (although complete target did)]
            best_hsp = max(hit, key=lambda hsp: hsp.bitscore)
            add_hmm_score(seq, variant_model, best_hsp, seq.variant_hmm ==
variant_model)
            except Exception as e:
            self.log.warning('Failed while loading scores for curated:
{}'.format(str(e)))
            pass
    # Search generic by our HMMs for histone types
    self.search_via_hmm(hmms_db=COMBINED_HMM_HISTTYPES_FILE,
out=CURATED_GEN_HISTVAR_RESULTS_FILE,
```

```
            sequences=CURATED_GENERIC_FASTA, procs=1, E=10)
    ##We have no any scores by variants for generic, but we can add scores to
histone types
    self.log.info("Loading hmmsearch for generic curated")
    for histtype_query in
SearchIO.parse(CURATED_GEN_HISTVAR_RESULTS_FILE,
"hmmer3-text"):
    histtype = histtype_query.id
    self.log.info("Loading hmmsearch for histone type: {}".format(histtype))
    for hit in histtype_query:
            accession = hit.id.split("|")[1]
            seq = Sequence.objects.get(id=accession)
            # best_hsp = max(hit, key=lambda hsp: hsp.bitscore)
            # hist_score = add_histone_score(seq,
Histone.objects.get(id=histtype), best_hsp)
            # add_generic_score(seq,
Variant.objects.get(id='generic_{}'.format(histtype)), hist_score)
            try:  # sometimes we get this:    [No individual domains that satisfy
reporting thresholds (although complete target did)]
            best_hsp = max(hit, key=lambda hsp: hsp.bitscore)
            hist_score = add_histone_score(seq, Histone.objects.get(id=histtype),
best_hsp)
            add_generic_score(seq,
Variant.objects.get(id='generic_{}'.format(histtype)), hist_score)
            except Exception as e:
            self.log.warning('Failed while loading scores for curated generic:
{}'.format(str(e)))
            pass
def classify_via_hmm(self, reset=True):
    """Classify loaded data in the histone database according to hmmer
```

```
results"""
    # TODO: Test method
    self.log.info("Classification of the data of HistoneDB...")
    # accessions = Sequence.objects.filter(reviewed=False).values_list('id',
flat=True)
    self.search_via_hmm(hmms_db=COMBINED_HMM_VARIANTS_FILE,
out=DB_HISTVARIANTS_HMM_RESULTS_FILE,
                        sequences=FULL_LENGTH_SEQS_FILE,
procs=HMMER_PROCS)
    for i in range(HMMER_PROCS):
    if not
os.path.isfile(DB_HISTVARIANTS_HMM_RESULTS_FILE.format('%02d' %
(i + 1))): continue
    self.log.info("Processing file %s ..." %
(DB_HISTVARIANTS_HMM_RESULTS_FILE.format('%02d' % (i + 1))))
self.load_hmm_classification_results(DB_HISTVARIANTS_HMM_RESULTS
_FILE.format('%02d' % (i + 1)))
    self.log.info('Total classified {} sequences'.format(
            Sequence.objects.exclude(variant_hmm__isnull=True).count()))
    self.load_generic_scores()
    self.canonical2H2AX()
def search_via_hmm(self, hmms_db, out, sequences, procs, E=10):
    """Use HMMs to search the nr database"""
    self.log.info("Searching HMMs...")
    self.log.info("Launching %d processes" % procs)
    child_procs = []
    for i in range(procs):
    try:
            self.log.info(" ".join(["nice", "hmmsearch", "-o", out.format('%02d' %
(i + 1)), "-E", str(E), "--notextw", hmms_db,sequences.format('%02d' % (i +
```

```
1))]))
            p = subprocess.Popen(["nice", "hmmsearch", "-o", out.format('%02d'
% (i + 1)), "-E", str(E), "--notextw", hmms_db,sequences.format('%02d' % (i +
1))])
    except TypeError:  # if procs=1 we have sequences as string without
formating %02d st the end
            self.log.info(" ".join(["nice", "hmmsearch", "-o", out, "-E", str(E),
"--notextw", hmms_db, sequences]))
            p = subprocess.Popen(["nice", "hmmsearch", "-o", out, "-E", str(E),
"--notextw", hmms_db, sequences])
    child_procs.append(p)
    for cp in child_procs:
    cp.wait()
# @transaction.atomic # looks like we cannot do it here, since transactions are
not atomic in this block
def load_hmm_classification_results(self, hmmerFile):
    """Save domain hits from a hmmer hmmsearch file into the Panchenko
Histone
    Variant DB format.
    Parameters:
    _____
    hmmerFile : string
    Path to HMMer hmmsearch output file.
    id_file : str
    Path to id file, to extract full lenght GIs
    """
    for variant_query in tqdm(SearchIO.parse(hmmerFile, "hmmer3-text")):
    self.log.info("Loading variant: {}".format(variant_query.id))
    variant_model = Variant.objects.get(id=variant_query.id)
    for hit in tqdm(variant_query):
```

```
            # Below we are fetching a list of headers if there are multiple headers
for identical sequences
            # Technically HUMMER might put the second and on gis in
description column.
            # The format should be strictly the genbank format: gi|343434|fafsf
gdgfdg gi|65656|534535 fdafaf
            headers = "{} {}".format(hit.id, hit.description).split('\x01')
            self.load_hmmhsps(headers, hit.hsps, variant_model)
def load_hmmhsps(self, headers, hsps, variant_model):
    ###Iterate through high scoring fragments.
    for hsp in hsps:
    # Compare bit scores
    hmmthreshold_passed = hsp.bitscore >= variant_model.hmmthreshold
    ##Iterate through headers of identical sequences.
    for header in headers:
            # to distinct accession from description and if accession is like
pir||S24178 get S24178
            accession = header.split(" ")[0]
            seqs = Sequence.objects.filter(id=accession)
            if len(seqs) <= 0:
            self.log.error("New sequence is found: {}. This is
strange.".format(accession))
            continue
            # Now if loaded bit score is greater than current, reassign variant and
update scores. Else, append score
            seq = seqs.first()
            if (seq.reviewed == True):
            continue  # we do not want to alter a reviewed sequence!
            if not hmmthreshold_passed:
            add_hmm_score(seq, variant_model, hsp,
```

```
best=hmmthreshold_passed)
                continue
                best_scores =
seq.all_model_hmm_scores.filter(used_for_classification=True)
                if len(best_scores) > 0:
                ##Sequence have passed the threshold for one of previous models.
                best_score = best_scores.first()
                if hsp.bitscore > best_score.score:
                # best scoring
                seq.variant_hmm = variant_model
                best_score_2 = ScoreHmm.objects.get(id=best_score.id)
                best_score_2.used_for_classification = False
                best_score_2.save()
                seq.save()
                add_hmm_score(seq, variant_model, hsp, best=True)
                else:
                add_hmm_score(seq, variant_model, hsp, best=False)
                else:
                # No previous model passed the threshold, it is the first
                seq.variant_hmm = variant_model
                seq.save()
                add_hmm_score(seq, variant_model, hsp, best=True)
def load_generic_scores(self):
    for hist_type in ['H2A', 'H2B', 'H3', 'H4', 'H1']:
    generic_model_sequences =
Sequence.objects.filter(histone_type__id=hist_type, variant_hmm__isnull=True)
    self.log.info("Found %d seqs did not pass threshold for %s" %
(generic_model_sequences.count(), hist_type))
    for generic_model_seq in generic_model_sequences:
                # self.log.error('variant_hmm is null?
```

```
{}'.format(generic_model_seq.variant_hmm))
                if generic_model_seq.reviewed:
                self.log.error('Among unclassified sequences found reviewed. This is
strange.')
                # self.log.error('Seq variant {}'.format(generic_model_seq.variant))
                # self.log.error('Seq variant_hmm
{}'.format(generic_model_seq.variant_hmm))
                # self.log.error('Seq histone_type
{}'.format(generic_model_seq.histone_type))
                # self.log.error('Seq reviewed
{}'.format(generic_model_seq.reviewed))
                continue
                generic_model =
get_or_create_unknown_variant(hist_type=hist_type)
                generic_model_seq.variant_hmm = generic_model
                generic_model_seq.save()
                try:
                add_generic_score(generic_model_seq, generic_model,
generic_model_seq.histone_model_scores.first())
                except Exception as e:
                self.log.warning('Failed: {}'.format(str(e)))
                pass
        self.log.info("Total classified %d seqs as generic %s" %
(Sequence.objects.filter(variant_hmm__hist_type__id=hist_type,

variant_hmm__id__startswith='generic').count(), hist_type))
def canonical2H2AX(self):
        """Fix an issue where the canonical variant takes over sequence from
H2A.X.
        The H2A.X motif SQ[ED][YFL]$ is not strong enough, but is the correct
```

```
variant.
    """
    self.log.info('Starting canonical2H2AX transformation ...')
    for s in
Sequence.objects.filter(variant_hmm="canonical_H2A",reviewed=False,
sequence__regex="SQ[ED][YFLIA]$"):
    old_score = s.all_model_hmm_scores.get(used_for_classification=True)
    old_score.used_for_classification = False
    old_score.save()
    # new_score, created =
ScoreHmm.objects.get_or_create(variant_hmm__id="H2A.X",sequence=s)
    obj = ScoreHmm.objects.filter(variant_hmm__id="H2A.X",sequence=s)
    if(len(obj)>1):
        self.log.warning('More than one score object for one variant found -
strange!!!')
        self.log.warning(obj)
    if(len(obj)==0):
        new_score, created =
ScoreHmm.objects.get_or_create(variant_hmm__id="H2A.X",sequence=s)
    else:
        new_score=obj.first()
    new_score.used_for_classification = True
    new_score.regex = True
    s.variant_hmm_id="H2A.X"
    new_score.save()
    s.save()
# For BLAST classification
def get_scores_for_curated_via_blast(self):
    # Score.objects.filter(sequence__reviewed=True).delete()
    for sequence in Sequence.objects.filter(reviewed=True):
```

```
    seqs_file = os.path.join(settings.STATIC_ROOT_AUX, "browse", "blast",
"HistoneDB_sequences.fa")
    blastp = os.path.join(os.path.dirname(sys.executable), "blastp")
    # output = os.path.join("/", "tmp", "{}.xml".format(uuid.uuid4()))
    blastp_cline = NcbiblastpCommandline(
            cmd=blastp,
            db=seqs_file,
            evalue=.01, outfmt=5)
    result, error = blastp_cline(stdin="\n".join([s.format("fasta") for s in
[sequence]]))
    resultFile = io.BytesIO()
    resultFile.write(result.encode("utf-8"))
    resultFile.seek(0)
    for i, blast_record in enumerate(NCBIXML.parse(resultFile)):
            if len(blast_record.alignments) == 0:
            self.log.error('No BLAST record alignments for {} during adding
scores for curated sequences'.format(sequence.id))
            continue
            for algn in blast_record.alignments:
            algn_self = False if sequence.id != algn.hit_def.split("|")[0] else True
# alignment on itself
            for hsp in algn.hsps:
            add_score(sequence, sequence.variant, hsp, algn.hit_def.split("|")[0],
best=algn_self) ### looks like there is an error
def classify_via_blast(self, force=True):
    for hist_type in ['H1', 'H2A', 'H2B', 'H3', 'H4']:
    self.log.info("Predicting variants for {} via BLAST".format(hist_type))
    # sequences = [seq.format(format='fasta') for seq in
    #
Sequence.objects.filter(reviewed=False).filter(variant_hmm__hist_type__id=hist
```

```
_type)]
    sequences = [seq.format(format='fasta') for seq in

Sequence.objects.filter(reviewed=False).filter(histone_type__id=hist_type)]
    self.predict_variants_via_blast(sequences=sequences,
blastout=DB_HISTVARIANTS_BLAST_RESULTS_FILE,
                        blastdb=BLASTDB_FILE, hist_type=hist_type,

result_file=DB_HISTVARIANTS_PARSED_RESULTS_FILE,
                        do_search=force, load_to_db=True,
procs=BLAST_PROCS)
    # self.checkH2AX()
def predict_variants_via_blast(self, sequences, blastout, blastdb, hist_type,
result_file=None, do_search=True, load_to_db=True, procs=1):
    if do_search:
    self.search_blast(sequences=sequences, blastdb=blastdb.format(hist_type),
                blastout=blastout.format(hist_type, "%d"), procs=procs)
    for i in range(procs + 1):
    # if os.stat(blastout.format(hist_type, "%d") % i).st_size == 0: continue
    with open(blastout.format(hist_type, "%d") % i, 'r') as blastFile:
            if re.search(r'^\s*$', blastFile.read()):
            self.log.warning('Results of {} is
empty'.format(blastout.format(hist_type, "%d") % i))
            continue
    result =
self.parse_blast_search_out(blastFile_name=blastout.format(hist_type, "%d") %
i, hist_type=hist_type)
    if load_to_db:
            self.load_in_db(parsed_blastout=result, hist_type=hist_type)
    result = pd.DataFrame(result).fillna('').drop(columns=['hsp', 'hit_accession'])
```

```
    if result_file:
            result.to_csv(result_file.format(hist_type, "%d") % i, index=False)
            self.log.info("Predicted sequences saved to
{}".format(result_file.format(hist_type, '%d') % i))

def search_blast(self, sequences, blastdb, blastout, procs):
    self.log.info("Running BLASTP for {}
sequences...".format(len(sequences)))
    split_count = int(len(sequences)/procs)
    for i in range(procs+1):
    sequences_split = sequences[split_count * i:split_count * i + split_count]
    self.log.info('Starting Blast sequences for {}/{}'.format(i, procs))
    self.make_blastp(sequences_split, blastdb, save_to=blastout % i)
def make_blastp(self, sequences, blastdb, save_to):
    # log.error('Error:: sequences {}'.format(len(sequences)))
    # log.error('Error:: hasattr {}'.format(hasattr(sequences, '__iter__')))
    if not hasattr(sequences, '__iter__'):
    sequences = [sequences]
    blastp = os.path.join(os.path.dirname(sys.executable), "blastp")
    # output = os.path.join("/", "tmp", "{}.xml".format(uuid.uuid4()))
    blastp_cline = NcbiblastpCommandline(
    cmd=blastp,
    # db=os.path.join(settings.STATIC_ROOT_AUX, "browse", "blast",
"HistoneDB_sequences.fa"),
    db=blastdb,
    evalue=.01, outfmt=5)
    # evalue=0.004, outfmt=5)
    result, error = blastp_cline(stdin="\n".join([s.format("fasta") for s in
sequences]))
```

```
    with open(save_to, 'w') as f:
    f.write(result)
    log.info('Blast results saved to {}'.format(save_to))
def parse_blast_search_out(self, blastFile_name, hist_type):
    """Parse blastFile file and return results as dict.
    Parameters:
    ____________
    blastFile: blastFile
    return e.g.: [{'accession': NP_563627.1, 'histone_type': H3, 'histone_variant':
cenH3, 'score': 2.14, 'best': True,
                  'description': Histone superfamily protein [Arabidopsis
thaliana], 'hsp': best HSPObject, 'hit_accession': DAA13058.1},
                  {'accession': DAA13058.1, 'histone_type': H2B,
'histone_variant': H2B.W, 'score': 3.11, 'best': False,
                  'description': [Bos taurus], 'hsp': best HSPObject,
'hit_accession': DAA13058.1},]
    """
    blastFile = open(blastFile_name)
    result = []
    count_new = 0
    for i, blast_record in enumerate(NCBIXML.parse(blastFile)): # <Iteration>
    query_split = blast_record.query.split('|')
    # self.log.info('DEBUG:: query_split = {}'.format(query_split))
    accession, description = query_split[0], query_split[-1]
    if accession == 'pir' or accession == 'prf':
        accession = '|'.join(query_split[:3])
        self.log.info('Non-standard accession {} got from
{}'.format(accession, blast_record.query))
    if len(blast_record.alignments) == 0: # <Hit> count = 0
        # log.info("No blast hits for {} with e-value
```

```
{}".format(blast_record.query, blast_record.descriptions[0]))
            self.log.info("No blast hits for {}".format(blast_record.query))
            # raise InvalidFASTA("No blast hits for
{}.".format(blast_record.query))
            result.append({'accession': accession, 'histone_type': hist_type,
                'histone_variant': 'generic_{}'.format(hist_type),
                'score': .000001, 'best': True, 'description': description,
                'hsp': None, 'hit_accession': None})
            continue
    best_alignments = []
    for alignment in blast_record.alignments: # <Hit>
            best_algn_hsp = self.get_best_hsp(alignment.hsps, align_longer=.25 *
blast_record.query_letters)
            best_alignments.append({'hit_accession':
alignment.hit_def.split("|")[0],
                        'hit_variant': alignment.hit_def.split("|")[2],
                        'best_hsp': best_algn_hsp,
                        'score': best_algn_hsp.score})
    best_alignments = sorted(best_alignments, key=lambda algn: algn['score'],
reverse=True)
    if best_alignments[0]['hit_variant'] == 'macroH2A':
            f = self.check_features_macroH2A(query_accession=accession,

hsp_hit_start=best_alignments[0]['best_hsp'].sbjct_start,
    hsp_hit_end=best_alignments[0]['best_hsp'].sbjct_end)
            if not f: continue
    histone_variant = best_alignments[0]['hit_variant']
    result.append({'accession': accession, 'histone_type': hist_type,
'histone_variant': histone_variant,
                'score': best_alignments[0]['score'], 'best': True, 'description':
```

```
description,
                    'hsp': best_alignments[0]['best_hsp'], 'hit_accession':
best_alignments[0]['hit_accession']})
    blastFile.close()
    return result
def load_in_db(self, parsed_blastout, hist_type):
    """Save parsed_blastout given as list of dicts to HistoneDB for current
hist_type.
    Parameters:
    ____________
    parsed_blastout: list of dists
    e.g.: [{'accession': NP_563627.1, 'histone_type': H3, 'histone_variant':
cenH3, 'score': 2.14, 'best': True,
        'description': Histone superfamily protein [Arabidopsis thaliana], 'hsp':
best HSPObject},
        {'accession': DAA13058.1, 'histone_type': H2B, 'histone_variant':
H2B.W, 'score': 3.11, 'best': False,
        'description': [Bos taurus], 'hsp': best HSPObject},]
    hist_type: string of histone type id
    """
    self.log.info("Loading BLASTP data for {} into
HistoneDB...".format(hist_type))
    for record in parsed_blastout:
    seq = Sequence.objects.get(id=record['accession'])
    if not seq: self.log.error('There is no such sequence {} in database. This is
strange.'.format(seq.id))
    variant_model = Variant.objects.get(id=record['histone_variant'])
    if record['best']:
        seq.variant = variant_model
        seq.save()
```

```
    try:
        if record['hit_accession']: add_score(seq, variant_model, record['hsp'],
record['hit_accession'], best=True)
        else: add_score(seq, variant_model)
    except Exception as e:
        self.log.warning('Failed to add histone variant score:
{}'.format(str(e)))
        pass
    self.log.info('Classified {} from {} in
database'.format(Sequence.objects.exclude(variant=None).count(),
                                        Sequence.objects.all().count()))
    non_classified = Sequence.objects.filter(variant=None,
histone_type__id=hist_type)
    self.log.info('Found {} sequences are not classified for
{}'.format(non_classified.count(), hist_type))
    for s in non_classified:
    s.variant = Variant.objects.get(id='generic_{}'.format(s.histone_type.id))
    s.save()
    add_score(s, s.variant)
    self.log.info('All these sequences classified as generic_{}'.format(hist_type))
def get_best_hsp(self, hsps, align_longer=0):
    best_alignment_hsp = hsps[0]
    for hsp in hsps[1:]:
    if hsp.score > best_alignment_hsp.score and hsp.align_length >
align_longer:
        best_alignment_hsp = hsp
    return best_alignment_hsp
def check_features_macroH2A(self, query_accession, hsp_hit_start,
hsp_hit_end):
    feature = Feature.objects.filter(template__variant='macroH2A',
```

```
name='Macro domain').first()
    ratio = (min(hsp_hit_end, feature.end) - max(hsp_hit_start, feature.start)) /
(feature.end - feature.start)
    self.log.info(
    '{} expected as macroH2A with ratio={} of macro domain contained in
hsp'.format(query_accession, ratio))
    if ratio < .8:
    self.log.info('{} expected as macroH2A cannot pass 0.8
ratio_treshhold'.format(query_accession))
    return False
    return True
def checkH2AX(self):
    """Fix an issue where the canonical variant takes over sequence from
H2A.X.
    The H2A.X motif SQ[ED][YFL]$ is not strong enough, but is the correct
variant.
    """
    self.log.info('Starting canonical2H2AX transformation ...')
    seqs =
Sequence.objects.filter(variant__id="canonical_H2A",reviewed=False,
sequence__regex="SQ[ED][YFLIA]$")
    self.log.info('Found {} sequences classified as canonical H2A with
H2A.X-motif'.format(seqs.count()))
    for s in seqs:
    old_score = s.all_model_scores.get(used_for_classification=True)
    old_score.used_for_classification = False
    old_score.save()
    obj = Score.objects.filter(variant__id="H2A.X",sequence=s)
    if(len(obj)>1):
            self.log.warning('More than one score object for one variant found -
```

```
strange!!!')
            self.log.warning(obj)
        if(len(obj)==0):
            # new_score, created =
Score.objects.get_or_create(variant__id="H2A.X",sequence=s)
            add_score(s, Variant.objects.get(id="H2A.X"))
        else:
            new_score=obj.first()
            new_score.used_for_classification = True
            # new_score.regex = True
            new_score.save()
        s.variant_id="H2A.X"
        s.save()
        self.log.info('Starting H2AX2generic transformation ...')
        seqs = Sequence.objects.filter(variant__id="H2A.X",
reviewed=False).exclude(sequence__regex="SQ[ED][YFLIA]$")
        self.log.info('Found {} sequences classified as H2A.X without
H2A.X-motif'.format(seqs.count()))
        for s in seqs:
        old_score = s.all_model_scores.get(used_for_classification=True)
        old_score.used_for_classification = False
        old_score.save()
        s.variant_id="generic_H2A"
        s.save()
        add_score(s, s.variant)
# Statistics
def get_stats_hmm(self):
        get_stats(start_time=self.start_time, filename='classifyvariants_hmm')
def get_stats_blast(self):
        get_stats(start_time=self.start_time, filename='classifyvariants_blast')
```