Федеральное государственное автономное образовательное учреждение   
высшего образования   
«Российский национальный исследовательский медицинский университет имени Н.И. Пирогова»   
Министерства здравоохранения Российской Федерации   
(ФГБОУ ВО РНИМУ им. Н.И. Пирогова Минздрава России)

Медико-биологический факультет  
Кафедра медицинской кибернетики и информатики им. С.А. Гаспаряна

Дипломная работа  
по специальности 30.05.03 «Медицинская кибернетика»

Смирнов Антон Сергеевич

Тема: Предсказание антигенов по

последовательностям белка

**Работа выполнена:**на кафедре биоинформатики МБФ  
ФГАОУ ВО РНИМУ   
им. Н.И. Пирогова

**Научный руководитель:**Лагунин А.А., д.б.н., профессор РАН,  
зав. кафедрой биоинформатики МБФ

Москва, 2023 г.

**Оглавление**

[Список сокращений 3](#_Toc136219684)

[Введение 5](#_Toc136219685)

[Глава 1. Литературный обзор 9](#_Toc136219686)

[1.1. Главный комплекс гистосовместимости 9](#_Toc136219687)

[1.2. Процессинг антигенов для MHC I 13](#_Toc136219688)

[1.3. Программы и сервисы для предсказания Т-клеточных эпитопов 20](#_Toc136219689)

[Глава 2. Материалы и методы 24](#_Toc136219690)

[2.1 Источники данных для моделирования 24](#_Toc136219691)

[2.1.1. Модель прогноза сайтов расщепления протеасомой 24](#_Toc136219692)

[2.1.2 Модель прогноза активности к ТАР 25](#_Toc136219693)

[2.1.3 Модель прогноза связывания с HLA 26](#_Toc136219694)

[2.2. Программа MultiPASS 26](#_Toc136219695)

[2.3. Построение моделей «структура-активность» 29](#_Toc136219696)

[2.4. Оценка качества моделей 30](#_Toc136219697)

[2.5. Веб-сервис 32](#_Toc136219698)

[2.6. Программное обеспечение для разработки 33](#_Toc136219699)

[Глава 3. Результаты 34](#_Toc136219700)

[3.1. Моделирование протеасомы 35](#_Toc136219701)

[3.2 Моделирование TAP 49](#_Toc136219702)

[3.3 Моделирование взаимодействия с HLA I 52](#_Toc136219703)

[3.4. Веб-сервис 55](#_Toc136219704)

[3.5. Обсуждение 59](#_Toc136219705)

[Глава 4. Заключение 62](#_Toc136219706)

[Глава 5. Выводы 64](#_Toc136219707)

[Список литературы 65](#_Toc136219708)

[Приложение 1 74](#_Toc136219709)

[Приложение 2 76](#_Toc136219710)

**Ошибка! Недопустимый объект гиперссылки.Ошибка! Недопустимый объект гиперссылки.Ошибка! Недопустимый объект гиперссылки.Ошибка! Недопустимый объект гиперссылки.Ошибка! Недопустимый объект гиперссылки.Ошибка! Недопустимый объект гиперссылки.Ошибка! Недопустимый объект гиперссылки.Ошибка! Недопустимый объект гиперссылки.Ошибка! Недопустимый объект гиперссылки.Ошибка! Недопустимый объект гиперссылки.Ошибка! Недопустимый объект гиперссылки.Ошибка! Недопустимый объект гиперссылки.Ошибка! Недопустимый объект гиперссылки.Ошибка! Недопустимый объект гиперссылки.Ошибка! Недопустимый объект гиперссылки.Ошибка! Недопустимый объект гиперссылки.Ошибка! Недопустимый объект гиперссылки.Ошибка! Недопустимый объект гиперссылки.Ошибка! Недопустимый объект гиперссылки.Ошибка! Недопустимый объект гиперссылки.Ошибка! Недопустимый объект гиперссылки.Ошибка! Недопустимый объект гиперссылки.Ошибка! Недопустимый объект гиперссылки.Ошибка! Недопустимый объект гиперссылки.Ошибка! Недопустимый объект гиперссылки.Ошибка! Недопустимый объект гиперссылки.Ошибка! Недопустимый объект гиперссылки.**

# Список сокращений

* ANN – artificial neural network – искусственная нейронная сеть
* AUROC (AUC) – Area Under Receiver Operation Curve – площадь под операционной кривой приемника
* CD - cluster of differntation - кластер дифференцировки
* CDR - complementarity determining region - регион, определяющий комплементарность
* CV – cross-validation – кросс-валидация
* DRiP - defective ribosomal products - дефективные рибосомальные продукты
* ERAAP - endoplasmatic reticulum aminopeptidase associated with antigen processing - аминопептидазы, ассоциированные с антигенным процессингом
* FTP – File Transfer Protocol – протокол передачи файлов
* HLA - Human Leukocytes Antigen - человеческий лейкоцитарный антиген, название человеческих генов МНС.
* HMM – hidden Markov models – скрытые Марковские модели
* HTTPS – Hyper Text Transfer Protocol Secure – защищенный протокол передачи гипертекстовых документов
* IAP – Invariant Accuracy of Prediction – Инвариантная точность предсказания
* IC50 - концентрации полумаксимального ингибирования
* IFN-γ – Интерферон-гамма
* IPv4 – Internet Protocol version 4 – интернет-протокол четвертой версии
* MCC – Mattew correlation coefficient – коэффициент корреляции Мэтью
* МНС - Major histocompatibility complex - главный комплекс гистосовместимости
* MNA - Multilevel Neighborhoods of Atoms – многоуровневые атомные окрестности
* PCPS - proteosome-catalyzed peptide splicing - пептидный сплайсинг, катализируемый протеосомой
* PLC – peptide-loading complex – пептид-загружающий комплекс
* SDF – Structure Data File – формат описания свойств химических соединений
* SSH – Secure Shell – протокол безопасной среды выполнения
* SVM – support vector machine – метод опорных векторов
* TAP – Transporter associated with Antigen Processing – транспортер, ассоциированный с антигенным процессингом
* UPS - ubiquitin-proteasomal system - убиквитин-протеасомальная система
* VPN – Virtual Private Network – виртуальная частная сеть
* VPS – Virtual Private Server – виртуальный частный сервер
* АПК - антигенпрезентирующая клетка
* БД – база данных
* мРНК - матричная рибонуклеиновая кислота
* нМ – наномоль/литр
* СУБД – система управления базой данных
* ТКР - Т-клеточный рецептор
* ЦПМ - цитоплазматическая мембран
* ЭПР - эндоплазматический ретикулюм

# Введение

Каждый живой организм на нашей планете стремится передать свой генетический материал потомкам. Для осуществления этой цели у каждого вида организмов должны быть механизмы защиты от конкурентов. Эти механизмы эволюционируют вместе с организмами. Примерно 500 миллионов лет назад, у хордовых возник ароморфоз: адаптивная иммунная система [1]. Иммунная система - специализированная система органов и тканей, обеспечивающих иммунитет [2]. Иммунитет - способ защиты организма от живых тел и веществ, несущих признаки чужеродной информации [2]. Вещества, которые несут такие признаки называются антигенами.

Адаптивная иммунная система названа так потому, что спектр антигенов, против которых защищает адаптивная иммунная система, не наследуется, а формируется в процессе жизни организма и определяется антигенами, встреченными организмом на протяжении жизни. То есть иммунная система организма адаптируется к своему окружению. Т- и В-лимфоциты, основные составляющие адаптивной иммунной системы, распознают участок молекулы антигена, которые называется эпитоп (антигенная детерминанта) с помощью специализированных молекул-рецепторов на своих цитоплазматических мембранах.

Существуют разница в том, как и с помощью каких молекул распознается антиген Т- и В-лимфоцитами. B-лимфоциты способны распознавать нативный антиген с помощью своего В-клеточного рецептора. Т-лимфоциты способны распознавать только антигены, презентированные на главном комплексе гистосовместимости (МНС), с помощью своих Т-клеточных рецепторов (ТКР). Такое ограничение называют в иммунологии МНС-рестрикцией. МНС представляет своеобразную систему “свой-чужой”, в состав которой входят не только молекулы, отвечающие за процессинг и презентацию антигенов, но и помогающие в реализации иммунного ответа. Для развития эффективного иммунного ответа, как правило, необходимо, чтобы Т-хелперы или Т-киллеры с помощью своих ТКР распознали антиген и передали сигнал о наличии чужеродного антигена. В данной работе будет рассмотрено, как организм подготавливает антигены для распознавания Т-лимфоцитами и как этот процесс можно предсказывать, и будет предложен новый подход к предсказанию результата этого процесса.

Математическая оценка возможного количества -цепей ТКР равно [3]. Количество возможных белков длины n из 20 основных протеиногенных аминокислот равняется и из этого разнообразия возможно появление ещё большего числа пептидов, которые могут быть эпитопами. Невозможно и нецелесообразно проверять все сочетания этого разнообразия на предмет наличия или отсутствия взаимодействия экспериментально. Однако за 50 лет активной работы был накоплен огромный массив данных о процессах подготовки антигенов, их презентации и распознавания целевыми молекулами, в нашем случае ТКР. Эти данные аккумулированы в различные базы данных и базы знаний, которые позволяют учёным строить математические модели для предсказания результатов разных этапов процессинга антигенов, наличия или отсутствия иммуногенности и аллергенности эпитопов.

Над созданием математических моделей прогноза эпитопов работает большое количество ученых уже больше двадцати лет: первые программы начали появляться в начале нулевых годов этого века. Данные модели крайне полезны в вакцинологии и онкологии. В вакцинологии они используются для проектирования вакцин от различных заболеваний, в частности от лихорадки Эбола и COVID-19 [4, 5]. Методология разработки вакцин, когда сначала ведут компьютерный поиск антигенов, предсказывают их иммунногенность, а потом проверяют экспериментально, называется обратной вакцинологией (reverse vaccinology). В онкологии они используются для предсказания неоантигенов. Неоантигены - новые антигены, которых нет в здоровых клетках и которые появились в раковых в процессе накопления новых генетических вариантов в сравнении с исходными, герминальными клетками. Точное предсказание антигенов и неоантигенов открывает новые возможности в персонализированной иммунотерапии опухолей [6]. С течением времени качество моделей заметно росло и современные модели показывают хорошие результаты, но они ограничены небольшой выборкой людей, к которым они применимы. Существующие аналоги, как правило, позволяют предсказывать результат одного из этапов процессинга и работают на алгоритмах, требующих эмбеддинга, то есть представления последовательности белка в виде числовых векторов. Самые современные программы работают на нейронных сетях различных архитектур. Их главная проблема в том, что крайне трудно определить, почему программа пришла к такому выводу.

**Цель** работы – создать подход на основе моделей связи “структура-активность” процессинга антигенов для МНС первого класса у человека (Homo sapiens) на основе структурных формул белков и их фрагментов. **Новизна** настоящей работы заключается в использовании представления белков в виде совокупности молекулярных фрагментов, а не числовых векторов. Кроме того, для построения некоторых моделей используются свежеопубликованные данные.

Для выполнения цели данной работы необходимо решить следующие **задачи**:

1. Подготовить выборки для обучения и валидации моделей прогноза сайтов расщепления белков протеасомой, активности белковых фрагментов по отношению к ТАР транспортеру, связывания белковых фрагментов с конкретными аллелями MHC I;
2. Используя подготовленные выборки построить модели связи структура-активность для прогноза сайтов расщепления белков протеасомой, активности белковых фрагментов по отношению к ТАР транспортеру, связывания белковых фрагментов с конкретными аллелями MHC I и оценить их качество;
3. Объединить построенные модели в единую систему для прогноза эпитопов, презентируемых на МНС I
4. Разработать прототип веб-сервиса для свободного доступа к разработанной системе. Целевая аудитория разработанного сервиса – ученые-иммунологи, разработчики вакцин и препараты для иммунотерапии.

# Глава 1. Литературный обзор

## 1.1. Главный комплекс гистосовместимости

Главный комплекс гистосовместимости - комплекс тесно связанных генетических локусов, расположенных на коротком плече 6-ой хромосомы (6р), а также их белковых продуктов, отвечающих за развитие иммунного ответа и синтез трансплантационных антигенов. Этот комплекс содержит больше 200 генов [7]. Эти гены отличаются высокой полиморфностью, кодоминантностью [8]. В базе данных IPD-IMGT/HLA на 27 февраля 2023 года содержится информация о 35820 человеческих аллелях HLA [9]. В популяциях по генам МНС наблюдается большой процент гетерозигот [10]. Сцепленные группы генов МНС называются гаплотипами. Функции генов МНС заключается в следующем (жирным выделены функции, обозначающие цель существования такой системы):

* **Обеспечение процессинга и презентации антигенных пептидов - индукторов и мишеней иммунного ответа;**
* Обеспечение взаимодействия клеток;
* Распознавание собственных, измененных собственных и чужеродных клеток; запуск и реализация иммунного ответа против носителей генетической информации;
* Поддержание иммунологической толерантности (в том числе во время беременности);
* Участие в селекции Т-лимфоцитов;
* **Создание генетического разнообразия и обеспечение выживаемости вида.**

Все гены локуса МНС поделены на три класса: первый, второй и третий. На [рисунке 1](#fig-hlamap) представлена генетическая карта генов МНС [11].

|  |
| --- |
| **Рисунок 1**. Генетическая карта генов HLA. |

К классическим генам относят три пары генов, кодирующих - и - цепи МНС второго класса: HLA-DR, HLA-DP и HLA-DQ - и три гена кодирующие -цепь МНС первого класса: HLA-A, HLA-B, HLA-C. Зрелые продукты этих генов располагаются на цитоплазматической мембране и непосредственно участвуют в презентации антигенов. У многих людей кластер HLA-DR содержит дополнительный ген -цепи, который может соединяться с DR-цепью [12]. Помимо этого существует большое количество генов “неклассических” МНС, которые по структуре напоминают гены МНС, но их продукты не презентируют пептиды обычным Т-лимфоцитам, но выполняют ряд других функций в клетке [12]. Например, MIC A, MIC B ответственны за обеспечение контакта с NK-клетками, Т-лимфоцитами, HLA-DM и HLA-DO участвуют в процессинге антигенов для HLA второго класса [13, 14], HLA-G подавляют способность цитотоксических лимфоцитов секретировать IFN- и стимулируют секрецию TGF-, что играет важную роль в обеспечении толерантности матери к антигенам плода [15]. В локусе генов МНС содержатся гены, чьи продукты принимают участие в процессинге антигенов для MHC первого класса (TAPBP, TAP1, TAP2, LMP1, LMP2). Выделяют особую группу генов МНС третьего класса, продукты которых не участвуют в презентации антигенов, но играют важную роль в обеспечении иммунитета [16]. Это компоненты комплемента (С4А, С4В, С2), белки теплового шока (HSPA1A, HSPA1B), белки семейства фактора некроза опухоли (LTB, LTA, TNF-).

Рассмотрим строение главных комплексов гистосовместимости первого и второго класса. МНС первого класса - гетеродимер, состоящий из большой -цепи, которой он заякорен в цитоплазматической мембране, и -микроглобулина. -цепь имеет 3 домена. - и -домены формируют особый сайт - пептидсвязывающую бороздку, которую иногда в литературе называют в честь учёной, впервые описавшей структуру МНС I и МНС II, по фамилии Бьоркман [17]. Ген -микроглобулина не относится к генам МНС и расположен на 15 хромосоме. -цепь связана с -цепью нековалентными взаимодействиями. Этот комплекс нестабилен, пока в пептидсвязывающей бороздке нет пептида. В эндоплазматическом ретикулуме (ЭПР) он стабилизируется шапероном кальнексином. Комплексы, которые презентируют слабоаффинные эпитопы, быстро диссоциируют на цитоплазматическую мембрану (ЦПМ), и цепи возвращаются обратно в ЭПР благодаря реакциям белкового разрушения, связанным с ЭПР (endoplasmatic reticulum-associated protein degradation, ERAD) [18]. Геометрия бороздки Бьоркман у МНС первого класса “закрытая”, поэтому в неё помещаются пептиды ограниченного размера, длиной 8-10 аминокислот [19]. МНС второго класса представляет из себя гетеродимер. Обе цепи, и , имеют по два домена и связаны между собой нековалентными связями. Пептидсвязывающую бороздку формируют и домены [20], и она, в отличии от МНС первого класса, “открытая”, поэтому в неё помещаются более длинные пептиды. Обычно длина пептидов, презентируемых на МНС второго класса, колеблется от 12 до 25 аминокислот [21]. Как и МНС первого класса, МНС второго класса нестабилен без эпитопа.

Структуру пептидсвязывающей бороздки у МНС первого класса можно разделить на 6 карманов, причем крайние карманы А и F отличаются консервативностью [22]. Поэтому для связывания с МНС первого класса важную роль играют N-концевые и C-концевые остатки, которые связываются с крайними карманами. От силы, характера нековалентных взаимодействий этих карманов с некоторыми остатками пептида, которые называют якорными, зависит спектр связываемых эпитопов этим МНС. В зависимости от аллеля МНС может быть дополнительные первичные якорные остатки. Между ними могут находиться аминокислотные остатки с сильно различающимися физико-химическим свойствами. У МНС второго класса якорные остатки могут быть расположены на различном удалении друг от друга. Кроме этого, пептид может выходить за пределы пептидсвязывающей бороздки, следовательно, в его связывании также участвуют и другие структуры белка [12].

МНС двух классов презентируют свой пептид разным клеткам: МНС первого класса презентирует пептиды, распознаваемые CD8+ лимфоцитами, известные также как Т-киллеры, а МНС второго класса - CD4+ лимфоцитам (Т-хелперы) [23]. МНС первого класса экспрессируется практически на всех ядросодержащих клетках организма [24]. Единственное исключение - клетки ворсинчатого трофобласта. Показано, что экспрессия классических аллелей МНС I на них приводит к иммуноопосредованному аборту [25]. МНС второго класса экспрессируются конституциально клетками иммунной системы, в особенности антигенпрезентирующими клетками: дендритными клетками, макрофагами, В-лимфоцитами, эпителиальными клетками тимуса, но их экспрессия может быть индуцирована IFN- на остальных клетках [26].

От локализации патогена зависит, на каком МНС будут презентироваться его антигены. Например, разрушенные белки цитозольных патогенов - вирусов, некоторых бактерий - транспортируются в ЭПР и там встраиваются в вновь синтезированные МНС первого класса и оттуда готовые комплексы доставляются на цитоплазматическую мембрану [27]. Такой каскад реакций называют непосредственным презентированием. При непосредственном презентировании белки разрушаются специальным белковым комплексом - протеасомой. Помимо непосредственного презентирования существует каскад реакции, когда экзогенные пептиды презентируются на МНС I. Существуют патогены, способные не напрямую поражать АПК. Существует специальный тип дендритных клеток, который презентирует пептиды, полученные фаголизосомальной системой, на МНС первого, а не второго класса, как это протекает в других АПК [28]. Такой каскад называется перекрестной презентацией, а активацию Т-клеток, с помощью этого каскада - кросс-праймингом. Рассмотрение перекрестного презентирования и презенитрования на МНС II выходит за рамки данные работы, поэтому далее речь пойдет о процессах, протекающих при непосредственной презентации антигена.

## 1.2. Процессинг антигенов для MHC I

Процессинг антигенов для МНС первого класса проходит в 6 стадий [29]:

1. Получение антигена (фагоцитозом, пиноцитозом и прочими способами),
2. Присоединение метки к белку в цитоплазме,
3. Гидролиз меченого белка протеасомой на отдельные пептиды,
4. Транспорт пептидов в ЭПР,
5. Встраивание пептида в белок МНС первого класса,
6. Транспорт комплекса на ЦПМ.

В качестве метки используется особый небольшой консервативный белок - убиквитин. Мечение осуществляется специальной убиквитин-протеасомальной системой (Ubiquitin-proteasomal system, UPS). Каскад реакций мечения описан в этой работе [30]. Сначала 1 молекула убиквитина ковалентно присоединяется к остатку цистеина убиквитин-активирующего фермента своим С-концевым остатком с затратой АТФ, затем он пересносится на убиквитин-конъюгирующий фермент. Следующим этапом убиквитин-конъюгирующий фермент формирует комплекс с убиквитинлигазой, с которой связан некоторый белковый субстрат. Убиквитинлигаза катализирует реакцию ковалентного связывания остатка лизина субстрата с С-концевым остатком глицина молекулы убиквитина. Дальнейший рост метки осуществялется убиквитинлигазой. Метка состоит из 4 молекул убиквитина, соединённых между собой ковалентной связью между лизинами в 48 положении одной молекулы и С-концевым остатком глицина другой [12].

Меченые белки поступают в особый белковый комплекс - протеасому. Существует множество типов протеасом, в зависимости от комбинации составляющих её каталитических субъединиц [31]. Стандартная человеческая 26S протеасома состоит из 3 частей: активной зоны - 20S “бочонка”, который состоит из 28 субъединиц, которые организованы в стопку из 4 колец: 2 -кольца в центре и 2 -кольца по краям, - и 2 регуляторных 19S частей, регулирующие доступ внутрь каталитической области, распознающие меченый субстрат [32]. Каталитическими являются лишь 1, 2 и 5 субъединицы -колец. Разные субъединицы обладают разным типом химической активности, но нет определенных правил, позволяющих сказать по какому аминокислотному остатку точно произойдет расщепление в определенной последовательности. Характеристика каталитических субъединиц приведена в [таблице 1](#tbl-proteasome). Существуют специальные каталитические субъединицы, которые экспрессируются при наличии IFN-, и они обладают несколько другим профилем активности, чем стандартные каталитические субъединицы. Кроме этого, есть специальная пятая субъединица, которая экспрессируется только в кортикальном слое тимуса эпителиальными клетками. По типу входящих в активную зону каталитических субъединиц выделяют конститутивные протеасомы, иммунопротеасомы и тимопротеасомы. Конститутивные протеасомы - те протеасомы, в состав которых входят только конститутивные субъединицы , и . Иммунопротеосомами называют протеосому, в которой есть , и субъединицы [33]. Тимопротеосомой же называют специфичный для эпителиальных клеток кортикального слоя тимуса белковый комплекс, в состав которого входят , и субъедиинцы [34]. Также существуют промежуточные формы иммунопротеосом, в которых произошла неполная замена конститутивных субъединиц на индуцированные.

**Таблица 1**. Характеристика каталитических субъединиц протеасом. Суффикс i означает индуцированный, суффикс t - тимус-специфичный. Каспазо-подобная активность означает, что гидролиз пептидной связи предпочтителен после кислотных остатков аспартата и глутамат, трипсин-подобная активность - после основных остатков лизина и аргинина, химотрипсин-подобная - после гидрофобных остатков лейцина, триптофана, тирозина и фенилаланина.

| **Субъединица** | **Общепринятое имя** | **Альтернативное имя** | **Ген** | **Предпочтительная активность** |
| --- | --- | --- | --- | --- |
| Бета тип-6 |  | Y, Delta, LMP19, Pre3 | PSMB6 | Каспазо-подобная |
| Бета тип-9 |  | Lmp2, RING12, MC7 | PSMB9 | Химотрипсин-подобная |
| Бета тип-7 |  | Z, MC14, Lmp9, Pup1 | PSMB7 | Трипсин-подобная |
| Бета тип-10 |  | MECL-1, Lmp10 | PSMB10 | Трипсин-подобная |
| Бета тип-5 |  | X, epsilon, Lmp17, MB1, Pre2, Doa3, Prg1 | PSMB5 | Химотрипсин-подобная |
| Бета тип-8 |  | Lmp7, RING10, Y2, C13 | PSMB8 | Химотрипсин-подобная |
| Бета тип-11 |  | Thymus-specific β5 | PSMB11 | Химотрипсин-подобная |

Анализ субстратной специфичности показывает, что конститутивная протеасома предпочитает выпускать пептиды с гидрофобными или основными остатками на С-конце, но это не означает, что не бывает других аминокислот на С-конце, к тому же у протеасом нет особых предпочтений по другим позициям в итоговом пептиде, а наблюдаемые частоты аминокислотных остатков зависят от строения регуляторных и каталитических субъединиц используемой протеасомы [35]. Индуцированные субъединицы обладают отличающимся профилем каталитической активности: субъединица предпочитает гидрофобные остатки с разветвленной цепью, в отличие от , предпочитающей кислотные остатки, обладает ещё большим сродством к гидрофобным остаткам, чем , а профиль активности и идентичен [33]. субъединица обладает пониженной химотрипсин-подобной активностью без влияния на каспазо-подобную и трипсин-подобную активности, что увеличивает долю пептидов с кислотным С-концевым остатком [36]. Тимопротеасома участвует в положительной селекции CD8+-лимфоцитов [37]. Для положительной селекции лимфоцитов важно, чтобы они распознали комплекс “свой пептид-МНС”, поэтому тимопротеасома генерирует пептиды более похожие на собственные, в то время как для иммунопротеасомы важно генерировать пептиды максимально непохожие на собственные для скорейшей реализации иммунного ответа.

С помощью убиквитин-протеасомальная система (ubiquitin-proteasomal system – UPS) клетка избавляется как от поврежденных или ненужных белков, так и от дефективных рибосомальных продуктов (defective ribosomal products, DRiP). DRiP представляют из себя пептиды, транслированные с интронов на неправильно сплайсированных мРНК, транслированные со сдвигом рамки считывания и неправильно уложенные белки, меченные убиквитином для их быстрого разрушения протеосомой [12]. Нарушения утилизации DRiP могут приводить к различным заболеваниям [38]. DRiP также служат важным источником антигенных пептидов для презентирования на МНС I [39].

В цитоплазме с продуктами гидролиза белков протеасомой могут произойти следующие события: утрата пост-трансляционных модификаций, пептидный сплайсинг, катализируемый протеосомой (proteasome-catalyzed peptide splicing, PCPS), дальнейший гидролиз с N-конца цитозольными протеазами. Например, гликопротеины теряют N-связанные гликаны [40]. Эта реакция конвертирует остаток аспарагина в остаток аспартата, что важно для распознания некоторых опухолевых антигенов [41]. PCPS впервые наблюдали при анализе антигенов рака почки [42]. Несмотря на то, что некоторые ученые сомневаются в существовании этого явления [43], оно наблюдается и изучается разными коллективами учёных [44–48], и большая часть иммунопептидома, который презентируется на HLA I класса, до 35% в некоторых клеточных линиях, может быть представлен сплайсированными пептидами [49]. PCPS - процесс соединения двух пептидов, продуцированных протеасомой. Она же и катализирует реакцию сплайсинга. Механизм реакции заключается в атаке свободной аминогруппой одного пептида ацил-ферментного интермедиата [50]. Присоединение возможно в трех вариантах в зависимости от происхождения пептидов и порядка присоединения. Когда два пептида происходят из одного белка, говорят о цис-сплайсинге, если из двух разных белков, то о транс-сплайсинге. При цис-сплайсинге возможно два варианта. В первом случае пептиды присоединяются в том же порядке, как они идут в исходной последовательности, во втором случае же в обратном: пептид, который был ближе к С-концу исходного белка формирует N-конец сплайсированного пептида.

После выхода из протеасомы, пептиды достаточно длинные, чтобы не помещаться в пептидсвязывающую бороздку, могут дальше разрушаться с N-конца цитозольными протеазами. Есть убедительные доказательства, что после выхода из протеосомы, процессинг с С-конца не происходит в клетках [51, 52]. Показано, что участие в дальнейшем укорочении пептидов могут принимать участие пуромицин-чувствительная аминопептидаза, блеомуцингидролаза, анти-трипептидилпептидаза II [53, 54]. От полного разрушения пептиды защищены шаперонами, например, кольцевым комплексом (TCP-1 ring complex, TRiC) [55].

Далее пептиды подлежат транспорту в ЭПР, где должно произойти их встраивание в МНС I. Их транслокацию обеспечивают специальные транспортеры, ассоциированные с антигенным процессингом (transporters associated with antigen processing, TAP). Это два белка, ТАР-1 и ТАР-2, из семейства АТФ-связывающих кассет (ATP-binding cassette, ABC), состоящие из N-концевого трансмембраного домена и С-концевого нуклеотид-связывающего домена [56]. Гомодимеры этих белков не имеют функциональной активности [57]. Между трансмембранным доменом TAP-белков и белком тапасином есть солевой мостик, который важен для сборки пептид-загружающего комплекса (peptide-loading complex, PLC) [58]. Пептид связывается с транспортером, когда он в релаксированном состоянии, и инициирует в нем конформационные перестройки, в результате которых попадает в полость ЭПР [59]. Гидролиз АТФ возвращает транспортер в исходное состояние [60]. Пептиды не покидают ЭПР через TAP-транспортеры, так как их высокая концентрация в полости ЭПР вызывает транс-ингибирование комплекса [61]. Только один пептид в один момент времени связывается с ТАР-комплексом [62]. Оптимальная длина пептида для переноса 8-16 аминокислот, хотя ТАР-1/2 способен переносить пептиды до 40 аминокислот [63, 64]. Важными для связывания с транспортером являются первые три N-концевые и одна С-концевая аминокислота [65]. Есть некоторые предпочтения по якорным остаткам у ТАР-транспортеров, а между ними участки могут сильно отличаться по длине и физико-химическим свойствам [59]. В позиции 1 и 2 N-конца предпочтительны положительно заряженные аминокислоты, в позиции 3 предпочтительны ароматические аминокислоты, в первых трех позициях не должно быть пролина, гиброфобные или основные аминокислоты на С-конце [12, 66, 67].

После попадания пептида в ЭПР он сразу поступает в PLC. PLC состоит из собранной молекулы HLA I, шаперона кальретикулина, ТАР-ассоциированного белка тапасина и тиоловой редуктазы ERp57 [68]. Этот комплекс ответственен за присоединение эпитопа к молекуле HLA. Комплекс не загружает в молекулу HLA первый попавшийся пептид, а подбирает наиболее аффинный, удаляя низкоафинные. Этот процесс называется пептидным корректированием [12]. Кроме этого, в полости ЭПР есть специальные аминопептидазы, ассоциированные с антигенным процессингом (endoplasmatic reticulum aminopeptidase associated with antigen processing, ERAAP). Это два фермента, ERAP1 и ERAP2, которые представляют из цинковые протеазы с консервативным мотивом His-Glu-X-X-His-X18-Glu, которые также нарезают пептиды с N-конца [69]. Экспрессия этих ферментов увеличивается под действием IFN- [69]. Здоровые клетки с нокаутом этих генов продуцируют неиммунногенные пептиды или пептиды, которые вызывают гибель клетки из-за действия на них цитотоксических Т-лимфоцитов [70]. Пептиды, которым не удалось связаться с MHC удаляются из ЭПР с помощью АТФ-зависимого транспортера Sec61 [71].

Следует иметь в виду, что процессинг антигена не совсем последовательный процесс. Например, получение эпитопа IYMDGTADSFSF из фермента тирозиназы проходит следующие стадии [44]:

1. После синтеза белок попадает в ЭПР, где происходит N-гликозилирование по специфичным аспарагиновым остаткам;
2. После этого белок переносится в цитоплазму, но некоторые молекулы из них неправильно собираются, поэтому идут на утилизацию в UPS;
3. У неправильно собранных белков происходит потеря N-связанных гликанов с конверсией аспарагина в аспартат;
4. В протеасоме два пептида с такими конвертированными сайтами, IYMDGT и ADSFSF, соединяются (цис-прямой сплайсинг, катализированный протеасомой);
5. Пептид IYMDGTADSFSF переносится в ЭПР, загружается в молекулу MHC и полученный комплекс транспортируется на ЦПМ.

## 1.3. Программы и сервисы для предсказания Т-клеточных эпитопов

Моделированию антигенного процессинга чуть больше 20 лет и за это время были опробованы различные подходы и использовались разные источники данных. За это время накопился огромный массив данных, собранный независимо в несколько баз данных и баз знаний. Immune Epitope Database (IEDB) - наиболее значимый ресурс, который содержит, огромную базу данных иммунологических данных, онтологию и сервисы для предсказания различных взаимодействий [72]. Эта база данных (БД) - основа всех современных исследований в области моделирования связывания эпитопов с МНС, распознавания комплексов TCR-pMHC и других исследований, так как содержит в себе миллионы записей.

Также накоплен большой массив масс-спектрометрических данных, посвященных предсказанию продуктов расщепления белков протеасомой, который был объединен в БД InvitroSPI [73]. InvitroSPI объединяет три больших датасета *in vitro* экспериментов с протеасомой и содержит 16 631 уникальный пептид. Присутствуют как сплайсированные пептиды, так и несплайсированные. Однако есть данные, показывающие, что по *in vitro* данным нельзя прогнозировать, что будет *in vivo* [74]. До появления источников этой БД основным датасетом были данные о сайтах разрезания для дрожжевой енолазы I и -казеина [75, 76].

Данные по связыванию пептидов с TAP содержатся в двух БД: MHC BN 4.0 и AntiJen 2.0 [77, 78]. Однако на момент 09 апреля 2023 года выгрузить данные из AntiJen 2.0 не представлялось возможным. Также некоторые датасеты выложены в открытый доступ разработчиками программ [79]. Эти БД небольшие, насчитывают около тысячи записей. Активность по отношению к ТАР представлена в виде значений IC50 относительно пептида RRYNASTEL с радиоактивной меткой. Кроме IEDB существует ещё несколько небольших баз данных, хранящие порядка несколько десятков тысяч записей с данными об эпитопах, аллелях МНС и узнающих последовательностях ТКР: VDJdb, McPAS-TCR, EPIMHC, SYFPEITHI и другие [80–83].

Первые компьютерные программы, посвящённые моделированию отдельных шагов процессинга также появились в начале 2000-ых годов [84, 85]. Как было замечено, процессинг антигенов далеко не всегда протекает последовательно, однако уже в 2006 году была предпринята попытка его моделирования как единого процесса [86]. EpiJen моделирует три основополагающих этапа процессинга: расщепление белков протеасомой, транспорт в ЭПР и связывание с МНС - и выдает топ-5% пептидов, которые связываются с прогнозируемыми ТКР. Методы, ставшие классическими, собраны на сайте Immune Epitope Database Analysis Resource (IEDB-AR) [87]. Сервисы по предсказанию процессинга антигенов можно разделить на 2 группы: сервисы, которые предсказывают отдельные этапы процессинга, и сервисы, которые учитывают несколько этапов процессинга.

На данный момент исследователи моделируют реакцию разрушения белков протеасомой, транспорт пептидов ТАР, связывание пептидов с МНС. Наиболее известные примеры программ представлены в [таблице 2](#tbl-programs).

**Таблица 2**. Примеры программ для моделирования отдельных этапов процессинга антигенов

| Программа | Моделируемые этапы | Источники данных | Алгоритм | Литературный источник |
| --- | --- | --- | --- | --- |
| Pcleavage | Протеосома | Данные по енолазе и казеину, МНС BN, | SVM | [88] |
| NetChop | Протеосома | SYFPEITHI, MHCPEP, HIV Immunology Database, Данные по енолазе и казеину | ANN | [89] |
| PCPS | Протеосома | HIV Immunology Database, EPIMHC, Immuneepitope, Los Alamos database | N-граммы | [90] |
| PREDTAP | TAP | Экспериментальные данные | ANN + HMM | [91] |
| TAPREG | TAP | Экспериментальные данные, Antijen | SVM | [79] |
| NetMHCpan | MHC | IEDB | ANN | [92] |
| MHCSeqNet | MHC | IEDB, MHC IEDB, MHCflurry | ANN | [93] |

Программы NetCTLpan и MHCflurry моделируют несколько этапов процессинга. MHCflurry состоит из двух моделей на основе нейронных сетей: по предсказанию процессинга и связыванию с МНС - результат агрегирует с помощью логистической регрессии [94]. NetCTLpan предсказывает разрезание белков протеасомой, транспорт и связывание по отдельности [95]. Эти методы являются пан-аллельными, то есть предсказывают результат для почти всего многообразия известных аллелей HLA. Это достигается путем особого подхода представления аллелей в виде псевдопоследовательности остатков, находящихся на расстоянии до 4 Å от пептидов длиной 9 аминокислот в любых репрезентативных структурах комплекса HLA-пептид [92]. Активно ведется разработка пан-специфичных фреймворков для предсказания связывания пептидов с HLA и все они основаны на нейронных сетях [93, 96, 97]. Современные методы предсказания связывания пептида с МНС имеют высокий AUC ROC, больше 0,95, однако предсказания ограничены определенными аллелями MHC, а также какой-то конкретной длиной пептида, чаще в 9 аминокислот [98, 99]. Следует отметить, что в современных работах моделируют расщепление белков протеасомой и связывание пептидов с MHC. Активность по отношению к TAP транспортеру не является главным направлением моделирования из-за недостатка данных и трудоемкости экспериментов, в которых их получают.

# Глава 2. Материалы и методы

## 2.1 Источники данных для моделирования

Данные для моделирования были извлечены из открытых источников: опубликованных баз данных и статей из рецензируемых журналов. Отфильтрованные таблицы сохранялись в csv (comma-separated values) формате.

### 2.1.1. Модель прогноза сайтов расщепления протеасомой

Для моделей протеасомы существуют два типа данных: *in vitro* эксперименты и *in vivo* эксперименты. Эксперимент *in vitro* заключается в инкубировании небольших, длиной в 20-30 аминокислотных остатков, фрагментов белков с каталитическим ядром протеасомы или с цельной протеасомой в пробирке несколько часов и в определении получившихся пептидов методами масс-спектрометрии. Эксперимент *in vivo* заключается в элюировании комплексов пептид-МНС с поверхности ЦПМ живых клеток и определении аллеля МНС и последовательности пептида методами масс-спектрометрии. Для обучения и валидации моделей *in vitro* была использована БД InvitroSPI [73], дополненная обучающей выборкой программы pepsickle [100]. Были отобраны данные, полученные на 20S субъединице человеческой протеасомы с временем инкубации не более 5 часов; отфильтрованы сплайсированные пептиды, с посттрансляционными модификациями или ошибками химического синтеза, с ошибками записи последовательности. Правильно записанной считается последовательность, записанная в однобуквенном коде и состоящая из 20 основных протеиногенных аминокислот.

Данные *in vivo* были извлечены из копии IEDB, локально развернутой на базе СУБД MySQL. Для извлечения данных был написан скрипт на языке программирования SQL (таблица mhc\_elution). С его помощью были извлечены данные об экспериментах *in vivo* на клетках человека. Эксперименты in vivo заключается в выделении комплексов пептид-МНС из мембран клеток, элюировании пептида из комплекса и определении аллели MHC и последовательности пептида методами масс-спектрометрии. Критерии дальнейшей фильтрации:

* Пептид не связывается с HLA I
* Представлена неполная последовательность пептида
* Записи содержат следующие ключевые слова в комментариях: “bad”, “did not”, “TAP deficient”, “predict”, “ERAP1 silencing”
* Неправильно записанная последовательность пептида или белка, из которого происходит этот пептид
* Пептид короче 8 или длиннее 14 аминокислотных остатков
* Последовательность пептида встречается в референсном белке больше 1 раза или не встречается вовсе

После фильтрации из данных *in vivo* и *in vitro* были удалены повторяющиеся записи.

### 2.1.2 Модель прогноза активности к ТАР

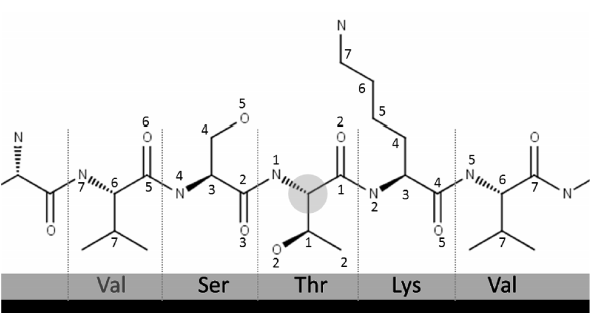
Для обучения и валидации были взяты данные из статьи про сервис TAPPRED [79]. Исходный набор данных скомпилирован из двух источников [77, 101] и значение меры активности – концентрации полумаксимального ингибирования (IC50) относительно пептида RRYNASTEL – был представлен в логарифмическом масштабе по основанию 2. Эти значения были измерены на микросомах, приготовленных из клеточной линии Sf9, где референсный пептид был помечен 125I по остатку тирозина[101]. Относительное IC50 изменяется от 0,1 до 338 нМ, что не позволяет отделить нам активные пептиды от неактивных, так как порог для неактивных пептидов, рекомендуемый в литературе равен 500 нМ [102]. Поэтому модель разделяет сильноактивные пептиды от слабоактивных и неактивных по порогу в 100 нМ, несколько большему принятому в литературе (50 нМ) [102]. Порог был увеличен для уменьшения доли ложно отрицательных результатов.

### 2.1.3 Модель прогноза связывания с HLA

Данные для обучения и валидации модели были взяты из IEDB и статьи про сервис MHCflurry [94]. Для извлечения данных из IEDB был написан SQL скрипт, аналогичный такому для извлечения данных для протеасомы (таблицы mhc\_bind и mhc\_elution). Данные, как и в случае с протеасомой, двух типов: *in vitro* и *in vivo*. Здесь эти данные можно объединить, так как и *in vitro*, и *in vivo* результаты определены в ходе непосредственного определения последовательностей пептидов и аллелей HLA. Для некоторых данных представлен качественный результат: активен или неактивен – для некоторых конкретное значение аффинности. Записи, для которых аффинность представлена в других единицах измерения, были отфильтрованы. Также были отфильтрованы записи, в комментариях к которым присутствовали ключевые слова “bad”, “did not”, “TAP deficient”, “predict”, “ERAP1 silencing”, “non-immunogenic”; содержащие неполные или неправильно записанные последовательности пептидов, неполно записанные HLA аллели, то есть записи, которые не позволяют однозначно идентифицировать последовательность α-цепи HLA I. Пептиды были разделены по аллель-специфичному порогу [103], если же такой не известен, то по рекомендуемому в литературе в 500 нМ [104]. Неактивные пептиды были отфильтрованы. Это связано с особенностью работы программы, используемой для моделирования.

## 2.2. Программа MultiPASS

Программа MultiPASS – коммерческая программа, разработанная в Институте биомедицинской химии им. В.Н. Ореховича. Является модификацией программы Prediction of Activity Spectra for Substances (PASS), разработанной там же, приспособленной для работы с биологическими последовательностями (ДНК, РНК, белки). Эта программа работает с дескрипторами многоуровневых атомных окрестностей (Multilevel Neighborhoods of Atoms, MNA). Эти дескрипторы являются подструктурами исходной молекулы. Генерация происходит по рекурсивному алгоритму: нулевой уровень – атомы в молекуле, первый уровень – дескрипторы нулевого уровня и атомы, связанные с дескриптором 0 уровня непосредственно ковалентной связью, второй уровень – дескрипторы первого уровня и атомы, связанные с атомами из дескриптора первого уровня ковалентной связью и так далее [105]. Наглядная схема представлена на рисунке 2.



**Рисунок 2**. Схема формирования MNA дескрипторов.

Допустим нам необходимы дескрипторы седьмого уровня. Атом, относительно которого идет формирование дескрипторов, отмечен полупрозрачным кругом. Нумерация обозначает порядок соседства для выбранного атома. Тогда дескрипторами седьмого уровня для выбранного атома будут являться все атомы, чем уровень соседства меньше или равен семи, в порядке связывания друг с другом. Процедура повторяется для каждого атома в молекуле. Для полученного множества уникальных MNA дескрипторов рассчитываются частотные оценки априорных вероятностей и правдоподобия ассоциации дескрипторов с некоторой активностью.

Такое множество дескрипторов учитывает тип атомов и порядок связей между ними, но не характер самих связей. MultiPASS основан на модифицированном наивном Байесовском классификаторе. Априорные вероятности и правдоподобия ассоциации дескрипторов с некоторой активностью оцениваются по входной выборке с использованием арксинус-преобразования Фишера [106]. При наличии нескольких классов в обучающей выборке MultiPASS берет за отрицательные примеры все примеры, которые не относятся к обучаемому классу. На выходе MultiPASS выдает таблицу, в которой для целевых веществ выставлен прогноз в диапазоне от минус одного до одного включительно. Минус единица означает, что программа уверена, что вещество не обладает этой активностью, единица означает уверенность программы в обладании данным веществом прогнозируемой активностью. MultiPASS работает только на операционной системе Windows.

Программа MultiPASS на вход принимает файлы в Structure-Data File (SDF) формате. SDF файл представляет из себя таблицу, в одной из колонок которой записаны структурные формулы веществ в формате MOL. Запись в файле выглядит следующим образом: сначала идет описание молекулы в MOL формате, далее идет описание свойств, запись заканчивается 4 символами доллара. MOL запись молекулы состоит из идентификатора, строки комментария, заголовка-описания молекулы, таблицы атомов и таблицы связей между атомами. Строка с названием свойств начинается с символа «больше». Название свойства заключено в угловые скобки, а его значение идет следующей строкой. Для использования MultiPASS был написан скрипт для конвертирования файлов с выборками из csv формата в sdf формат с помощью языка программирования Python, использующий библиотеку RDKit [107].

## 2.3. Построение моделей «структура-активность»

В основе построения моделей «структура-активность» лежит предположение, что некоторая биологическая активность химического соединения есть функция от её химической структуры [108]. Для построения моделей прогноза активности для TAP и HLA уже в сыром виде есть структура и связанная с ней активность. Для предсказания активности протеасомы к каждому остатку в белке необходимо получить структуру «окрестности» для каждого остатка. Построение модели прогноза сайтов расщепления белков протеасомой происходило по следующему алгоритму:

1. Поиск координат пептидов в их исходной последовательности
2. Для данных *in vitro* активными помечались остаток левее N-конца пептида и С-конец пептида, для данных *in vivo* только С-конец пептида. Все остальные остатки помечались как неактивные (*in vitro*) или случайный остаток из входящих в пептид, но не входящих в окно с активным остатком (*in vivo*).
3. Задается некоторый радиус окна w. В этой работе мы использовали радиус окна равный 2, 3, 4 аминокислотным остаткам.
4. От i = w до “длина последовательности – w - 1” включительно:
   1. Выбирается подпоследовательность s [i - w; i + w] (длина подпоследовательности всегда нечетное число)
   2. Присвоение подпоследовательности класса активный/неактивный в зависимости от класса центрального остатка
5. Крайним остаткам присваивается статус «неактивный»
6. Полученная таблица конвертируется в SDF формат
7. Построение модели в MultiPASS
8. Получение прогноза
9. Процедура восстановления последовательности из подпоследовательности
10. Оценка качества прогноза

Для моделей связывания с TAP и HLA процедура построения моделей проще: отфильтрованные выборки конвертируются в SDF формат и подаются на вход программе MultiPASS, после чего оценивается качество прогноза. Для построения моделей использовалась как консольная версия MultiPASS, так и версия с графическим интерфейсом.

## 2.4. Оценка качества моделей

Достоинство используемого подхода – малое число гиперпараметров. Для моделей ТАР и HLA он всего один – уровень MNA дескрипторов. Для модели протеасомы два – радиус используемого окна и уровень MNA дескрипторов. Подбор оптимальных гиперпараметров и оценка качества моделей осуществлялась с помощью процедуры пятикратной кросс-валидации.

K-fold кросс-валидация – процедура независимой оценки качества моделей, когда исходный набор разбивается k раз на обучающую и тестовую выборку так, чтобы тестовые наборы не пересекались [109]. Для каждого обучающего набора был сделан прогноз на тестовой выборке. Затем результаты прогноза объединялись в один файл, по которому рассчитывались метрики качества моделей. После оптимизации гиперпараметров для лучшей модели были рассчитаны следующие метрики помимо AUROC: чувствительность, специфичность, точность (accuracy), сбалансированная точность (balanced accuracy), средняя точность (average precision) и коэффициент корреляции Мэтью (MCC). Данные метрики рассчитываются по формулам 1-6 соответственно. Порог для расчета метрик рассчитывался по критерию Юдена [110]. Для разбиения на фолды и расчёта метрик были использованы пакеты caret[111] для языка программирования R и scikit-learn[112] для языка программирования Python.

где: TP – количество правильно предсказанных положительных случаев,

TN – количество правильно предсказанных отрицательных случаев,

FP – количество ложно предсказанных положительных случаев,

FN – количество ложно предсказанных отрицательных случаев,

Rn – чувствительность, рассчитанная для порога n

Pn – , рассчитанный для порога n

В качестве оптимизируемой при кросс-валидации метрики была выбрана площадь под операционной кривой приемника (Area Under Receiver Operation Curve, AUROC). AUROC рассчитывается по методу, описанному в следующей работе [113]. Вводится понятие инвариантной точности прогноза (Invariant Accuracy of Prediction, IAP), которое численно равно AUROC и рассчитывается по формуле 1,

где: N(C+) – количество положительных примеров,

N(С-) – количество отрицательных примеров,

N{P(C+) > P(C-)} – количество примеров, для которых оценка вероятности для случайно взятого положительного примера больше оценки вероятности для случайно взятого отрицательного примера.

Чем выше значение AUROC, тем выше качество модели, поэтому эта метрика подлежит максимизации.

Кроме пятикратной кросс-валидации использовалась внутренняя оценка качества модели – расчет AUROC по 20-кратной кросс-валидации (AUROC20CV) на каждом обучающем наборе, то есть обучающий набор делится на обучающую и тестовую выборки 20 раз, так чтобы получающиеся тестовые выборки не пересекались. AUROC20CV для каждого класса рассчитывает MultiPASS, используя формулу 7.

## 2.5. Веб-сервис

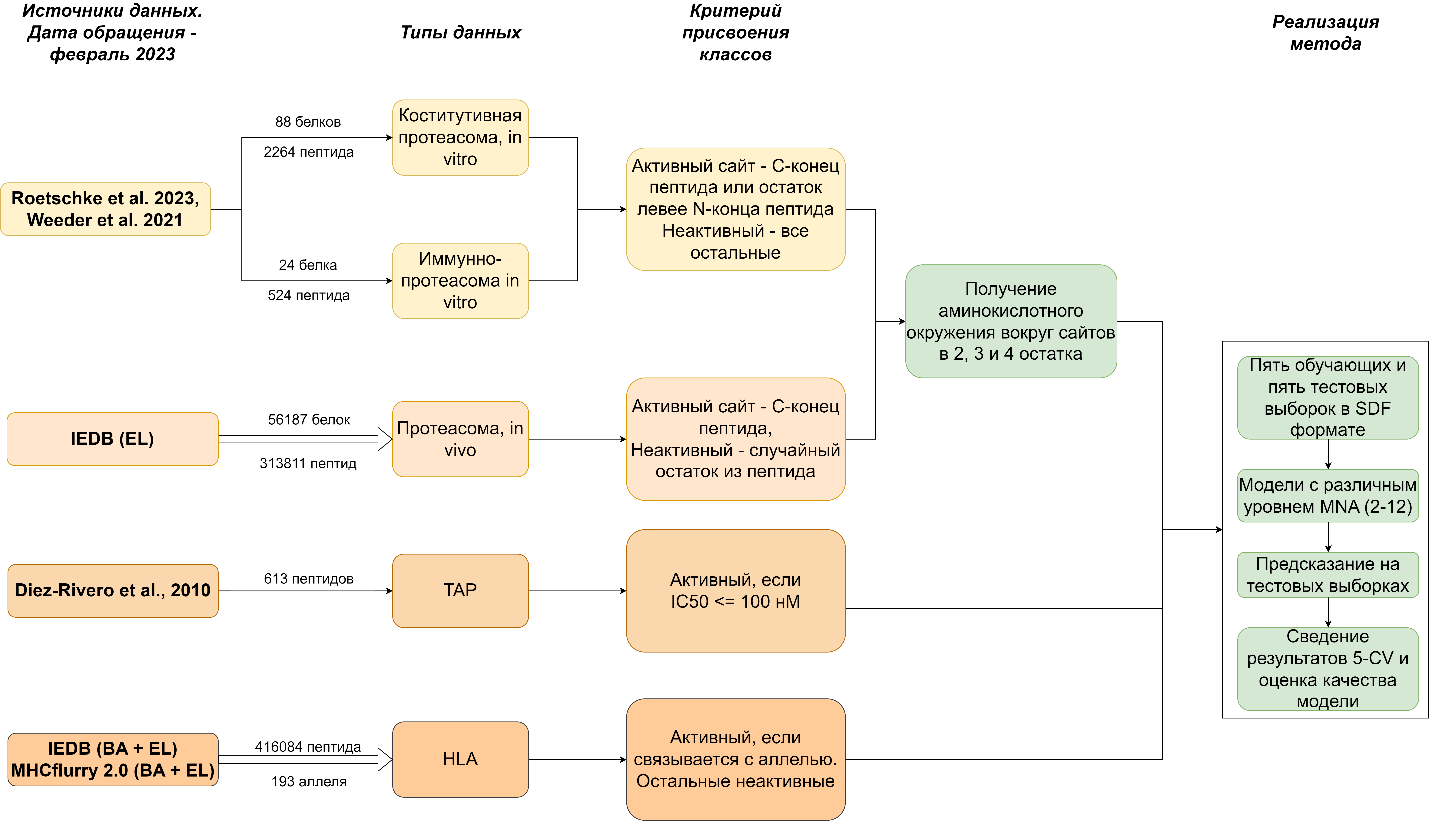
Веб-сервис разработан с помощью пакета shiny[114] для языка программирования R. Shiny Server был развернут на арендованном виртуальном частном сервере (Virtual Private Server, VPS) от российской компании Jino. VPS работает под операционной системой Ubuntu 20.04. MultiPASS не поддерживает Unix-подобные операционные системы, поэтому на VPS также был развернут VPN (Virtual Private Network, виртуальная частная сеть) сервер на основе стандарта OpenVPN, сервер по обмену данных по протоколу FTP (File Transfer Protocol, протокол обмена файлами) и сервер OpenSSH (SSH - Secure Shell, протокол безопасного выполнения команд на удаленном терминале). К VPN серверу подключена домашняя рабочая станция, на которой стоит консольная версия MultiPASS. Подключение через VPN необходимо для устойчивого соединения между сервером и станцией с известными внутри виртуальной сети статическими IPv4 адресами, по протоколу FTP сервер отправляет входные данные на станцию и получает от нее результат. Команды и скрипты промежуточной обработки данных, написанные на языке программирования Python, сервер запускает на станции через подключение по SSH.

## 2.6. Программное обеспечение для разработки

Все поставленные в этой работе задачи выполнялись с помощью языков программирования Python версии 3.9, R версии 4.2.3, SQL диалекта MySQL. Основные подзадачи, исполняемые веб-сервисом, написаны на языке программирования Python. Сам веб-сервис реализован с помощью фреймворка shiny на языке программирования R. Для работы с БД IEDB был развернут локальный MySQL Server версии 8.0, доступ к нему и работа с ним проводилась с помощью программы DataGrip от компании JetBrains. Для разработки на языках программирования Python, R использовалась программа PyCharm Professional от компании JetBrains. Разработка веб-сервиса осуществлялась также с помощью RStudio Server от компании Posit. Разработка велась в изолированном виртуальном окружении, созданном с помощью программы Anaconda. Веб-сервис работает на Shiny Server в связке с Nginx. Для связи между узлами веб-сервиса используются программы OpenVPN и vsftpd. В разработке использовалась система контроля версий git и удаленный репозиторий проектов GitHub. Для написания и запуска типовых процедур (пайплайнов) использовался Jupyter Notebook.

# Глава 3. Результаты

На рисунке 3 приведена краткая схема проводимого исследования. Она состоит из 2 частей: описания использованных данных и описания процесса моделирования. Данные по протеасоме было решено разделить на in vitro и in vivo, так как мы строим предположении о сайте расщепления по тому, что наблюдаем на молекуле MHC. То есть моделирование на данных in vivo строится на предположении, что антигенный процессинг после выхода пептида из протеасомы не идёт с С-конца, что, строго говоря, иногда не соблюдается и в клетке присутствует множество факторов, которые могут повлиять на конечный презентируемый продукт, в том числе неизвестные сегодня науке. Данные по TAP транспортеру представляют собой результат измерений in vitro. Данные по HLA было решено не разделять, так как они представляют собой результат непосредственного определения связывания пептидов с MHC различными методами. Так как это конечный этап антигенного процессинга, то нет промежуточных этапов, которые могли бы изменить последовательность презентируемого пептида, а варианты с посттрансляционными модификациями были отфильтрованы. Критерий присвоения классов были составлены с опорой на данные литературы и технические возможности программы MultiPASS. Следует отметить, что данных in vitro на порядок меньше, чем данных in vivo, так как эксперименты in vitro дороже и сложнее экспериментов in vivo: удешевлению экспериментов in vivo способствует широкое распространение высокопроизводительной масс-спектрометрии. Для эксперимента легче достать материал и сам эксперимент занимает меньше времени. Для эксперимента in vitro требуются очищенные белки, специальные, генномодифицированные клеточные линии и дорогие реактивы.



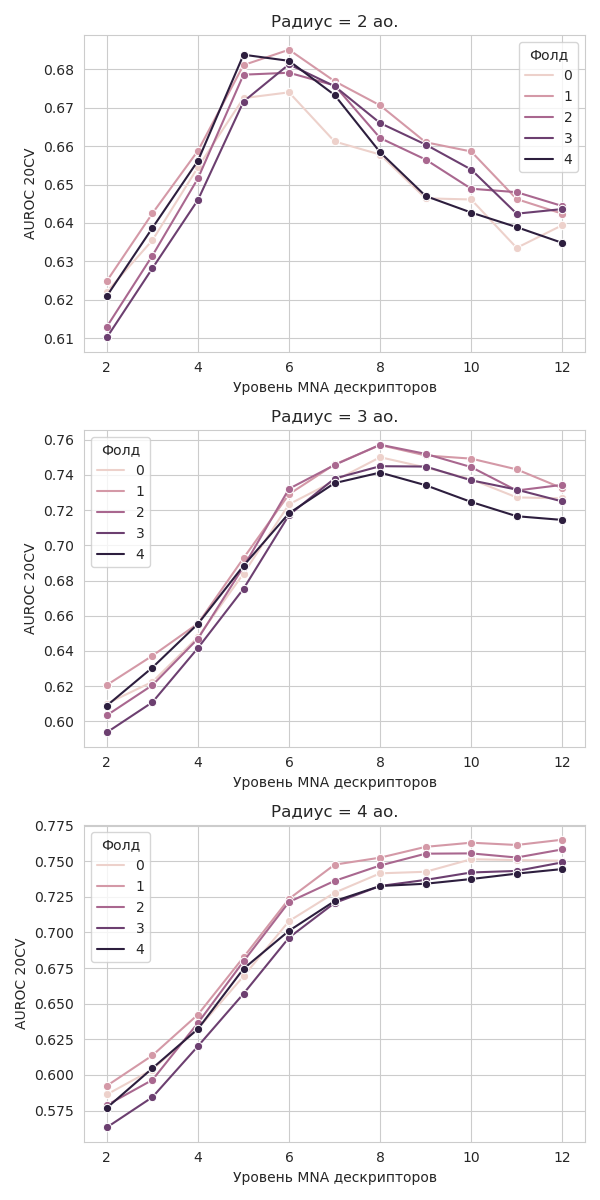
**Рисунок 3.** Схема проведенного исследования. BA – данные in vitro, EL – данные in vivo. Количество белков и пептидов приведено на момент после фильтрации и подготовки данных.

## 3.1. Моделирование протеасомы

Характеристика используемых наборов данных для моделей протеасомы представлена в Приложении 1. Для каждого набора данных были построены модели с разными значениями гиперпараметров. Радиус окна изменялся от двух до четырёх включительно с шагом в один, уровень MNA изменялся от двух до двенадцати включительно с шагом в один. Для пятикратной кросс-валидации набор данных разбивался пять раз, следовательно, на каждом наборе данных было построено 165 моделей.

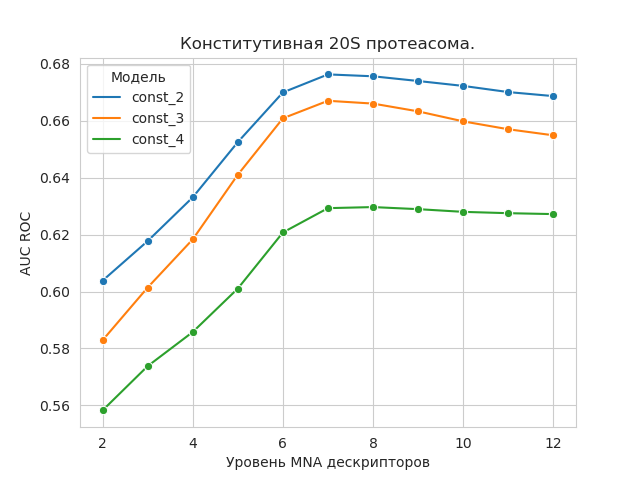
Внутренняя оценка моделей, построенных на данных in vitro для 20S конститутивной протеасомы представлена на рисунке 4. Для радиуса окна, равным двум, для каждого разбиения наблюдается типичный вид кривых: рост качества до некоторого оптимума, после которого наблюдается спад качества. Для радиуса окна, равным трём, наблюдается такой же вид кривых, но спад точности после оптимума не такой крутой. Для радиуса окна, равным четырём, кривая имеет сигмоидный вид: качество на высоких уровнях MNA дескрипторах выходит на плато. Разница между наибольшим и наименьшим значением точности растет пропорционально радиусу окна. С увеличением радиуса окна увеличивается уровень MNA дескриптора, при котором наблюдается максимальное качество. Кривые имеют одинаковый вид и по абсолютным значениям координат несильно друг от друга. По данным кривым можно сделать следующие промежуточные выводы:

1. Разбиение выборок при пятикратной кросс-валидации было сбалансировано;
2. Качество предсказания сайтов расщепления зависит от радиуса используемого окна прямо пропорционально;
3. Чем бо́льший радиус окна используется для моделирования, тем больший уровень MNA дескрипторов необходимо использовать для моделирования.



**Рисунок 4.** Зависимость метрики внутренней оценки качества моделей на данных in vitro 20S конститутивной протеасомы от уровня MNA дескрипторов при различном радиусе скользящего окна.

Внешняя оценка для моделей, построенных на данных in vitro 20S конститутивной протеасомы представлена на рисунке 5. По рисунку 5 можно заметить, что несмотря на прямую зависимость точности модели от радиуса окна при внутренней оценке качества модели, наилучшее качество имеют модели, построенные с использованием минимального радиуса окна вне зависимости от уровня MNA дескрипторов. Качество моделей при использовании больших уровнях MNA дескрипторов не изменяется или изменяется незначительно. Точка оптимума располагается на седьмом уровне MNA дескрипторов для каждого радиуса окна. Уровень качества наилучших моделей недостаточен, чтобы называть его удовлетворительном. Для практического использования такие модели непригодны.

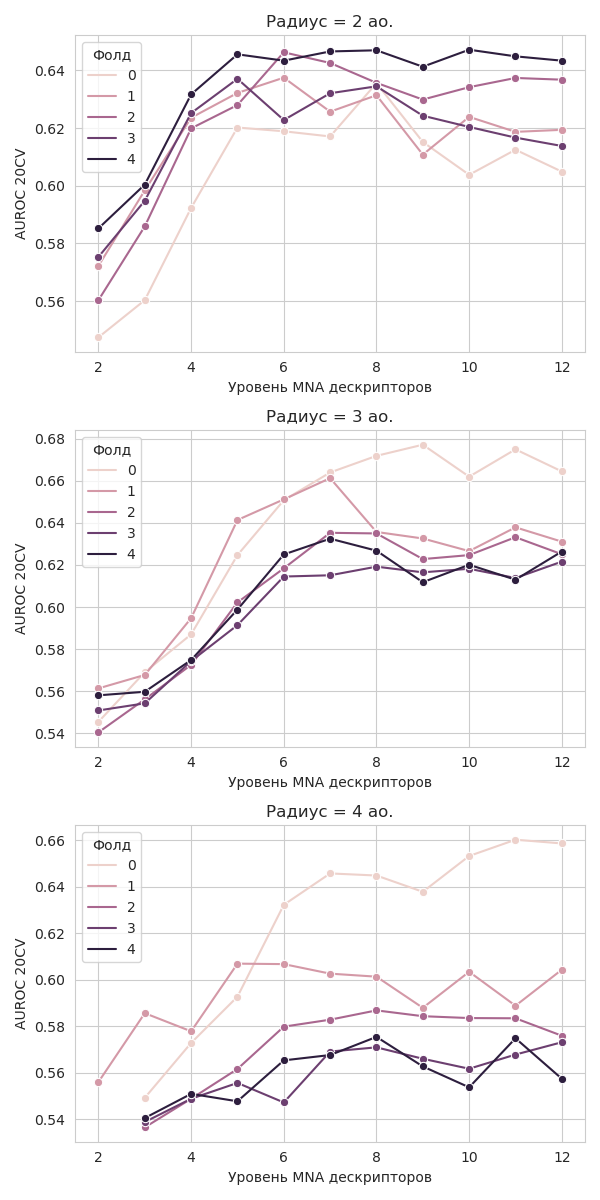


**Рисунок 5.** Зависимость внешней метрики качества моделей, построенных с использованием разного радиуса окна, от уровня MNA дескрипторов.

Внутренняя оценка моделей, построенных на данных in vitro для 20S иммуннопротеасомы представлена на рисунке 6. Можно заметить, что кривые могут значительно отличаться друг от друга разбросом и формой. Для радиуса окна, равным двум и трем, кривые визуально ещё укладываются в общую картину гиперболы, несмотря на выпадение кривых 0 для радиуса, равного двум, 0 и 1 для радиуса, равным трём из общей картины. Для радиуса, равного четырем, кривые имеют большой разброс друг относительно друга и различный друг от друга вид. Модели при радиусе окна в четыре и уровне MNA дескрипторов в два для разбиений с первого по четвертый не прошли внутренний контроль качества MultiPASS, поэтому не отображены на графике. По графикам внутреннего контроля качества тяжело судить об оптимальных значениях гиперпараметров. В отличие от моделей конститутивной протеасомы:

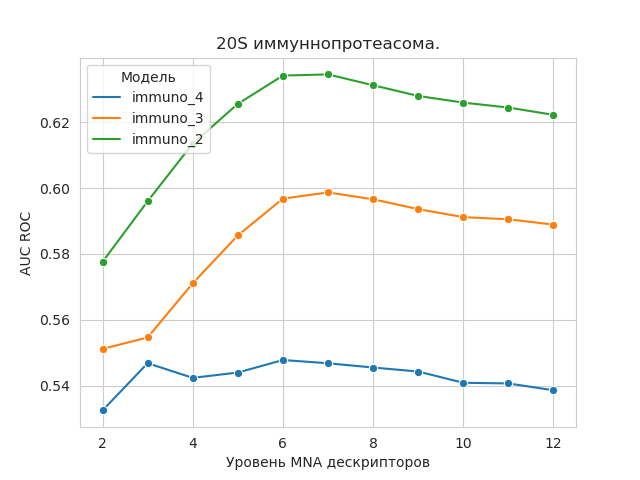
1. Разбиения отличаются крайней несбалансированностью;
2. Не наблюдается зависимости точности от радиуса окна и точности от радиуса окна и уровня MNA дескрипторов;
3. Для радиуса окна, равного четырем, крайне слабая зависимость точности прогноза от уровня MNA дескрипторов.

Качество моделей находится на неудовлетворительном уровне при всех выбранных вариантах значений гиперпараметров.



**Рисунок 6.** Зависимость метрики внутренней оценки качества моделей на данных in vitro 20S иммунопротеасомы от уровня MNA дескрипторов при различном радиусе скользящего окна.

Внешняя оценка для моделей, построенных на данных in vitro 20S иммуннопротеасомы представлена на рисунке 7.

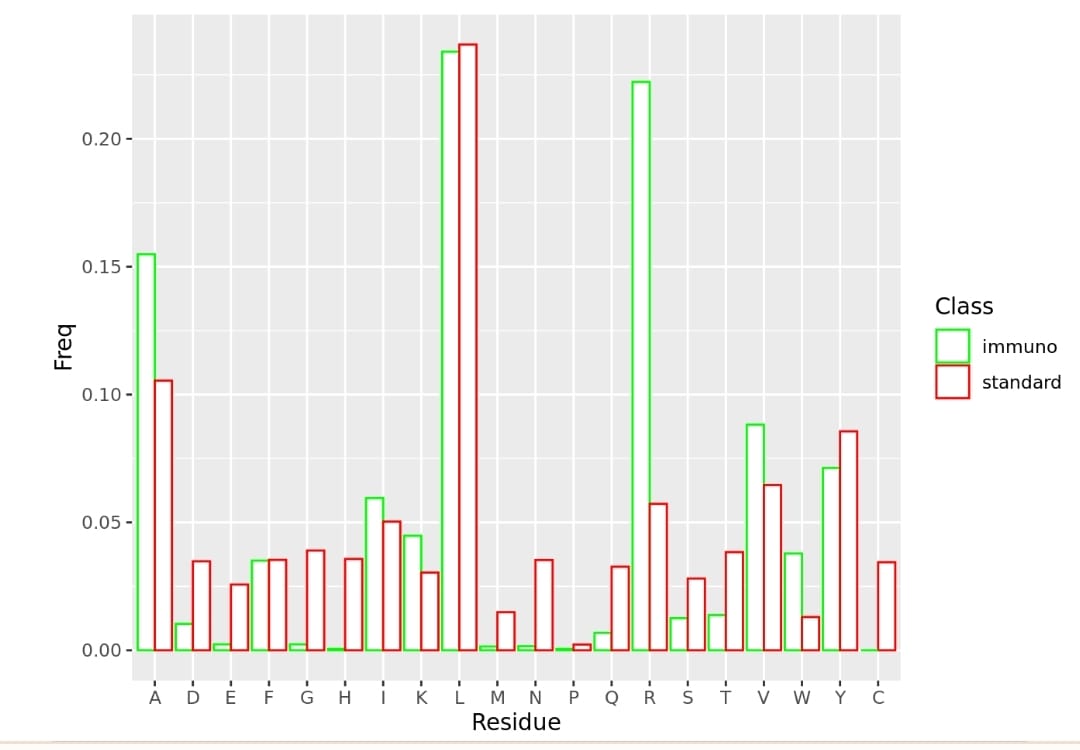


**Рисунок 7.** Зависимость внешней метрики качества моделей от уровня MNA дескрипторов.

Как и для конститутивной протеасомы наилучшее качество показали модели с использованием минимального окна. Для радиуса окна, равного двум и трем, оптимум MNA дескрипторов равен семи, как и для конститутивной протеасомы. Однако для радиуса окна, равного четырем, наблюдается отсутствие зависимости качества моделей от уровня MNA дескрипторов. Качество этих моделей находится практически на уровне случайного классификатора. Качество моделей с оптимальными значениями гиперпараметров также оставляет желать лучшего. Для практического использования такие модели непригодны.

Причина низкого качества моделей, построенных на данных in vitro может заключаться в самих данных и в принципах разметки. Отрицательные и положительные примеры могут располагаться друг от друга по соседству в исходной последовательности, могут чередоваться. Нередка ситуация, когда пептидные фрагменты является продуктами последовательного расщепления длинной последовательности. Например, исходная последовательность – FDSTRYPLKMVCWTLK, а продукты – FDSTRYPLKMV, FDSTRYPLK, – FDSTRY, тогда разметка активности с пятой по двенадцатую аминокислоту будет выглядеть следующим образом: 01001010. Таким образом, сайты могут быть разделены друг от друга всего лишь одним остатком и при использовании больших радиусов окна могут быть неправильно определены дескрипторы, ответственные за формирование сайта расщепления, поэтому нецелесообразно использовать бо́льший радиус окна. К тому же нецелесообразно увеличивать его из-за небольшой длины исходных белков, использованных для экспериментов, которые по сути своей являются частями некоторых больших белков. Использование радиуса окна, равного в единицу, было сочтено недостаточным, исходя из положений, что выявленные паттерны из трех остатков будут встречаться гораздо чаще, чем выявленные паттерны из пяти остатков, и приводит к кратному увеличению число ложно положительных результатов.

По рисунку 7 можно заметить, что профиль каталитической активности у конститутивной протеасомы и иммуннопротеасомы отличается. Например, иммуннопротеасома гораздо более активна по остатку аргинина, чем конститутивная протеасома, но неактивна по остатку цистеина и практически неактивна по остаткам метионина и аспарагина. Конститутивная протеасома проявляет одинаковую активность по отношению к большинству аминокислот, в отличие от иммуннопротеасомы. Использование только аминокислот невозможно, так как по данным литературы и проведенным исследованиям данных протеасомы работают по всем остаткам, но с разной частотой (рис. 7).

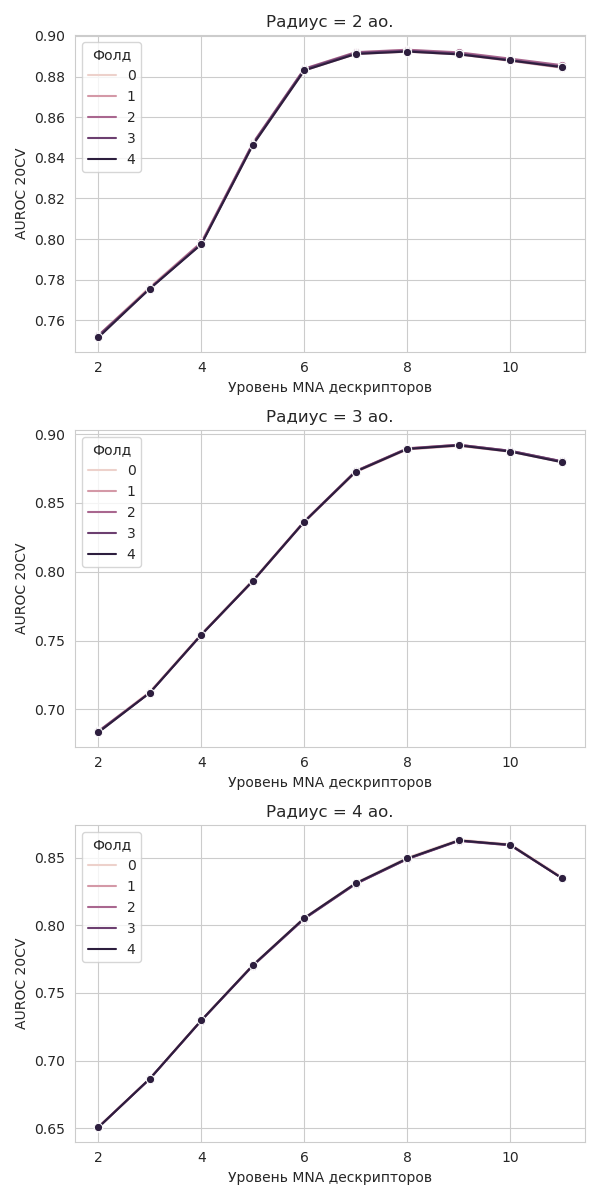


**Рисунок 8.** Относительная частота встречаемости аминокислотных остатков в сайтах расщепления белков протеасомой. Оценено по данным работы [115].

**#TODO – вставь гистограмму частот разрезания по остаткам**

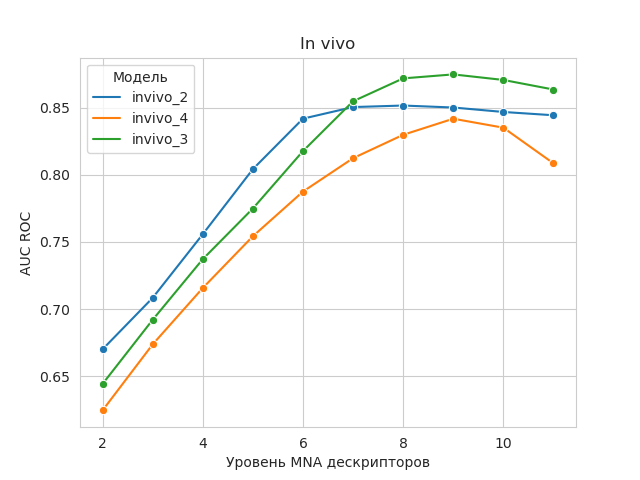
Внутренняя оценка моделей, построенных на данных in vivo представлена на рисунке 9. Кривые совпадают, так как датасет является практически идеально сбалансированным по количеству примеров в классах благодаря алгоритму формирования выборки и этих примеров довольно большое количество. На графиках не представлены данные для двенадцатого уровня MNA дескрипторов, так как модели получились такими же, как и при одиннадцатом из-за технических особенностей программы MultiPASS, о которых будет сказано ниже. Существуют некоторые отличия от результатов, полученных на конститутивной протеасоме. Заметно, что максимально достигаемое качество моделей при использовании радиуса окна в четыре меньше, чем при радиусе окна, равного двум и трём. Разница между минимальным и максимальным значением точности не зависит от радиуса окна. Следует отметить, что уровень MNA дескрипторов, при котором достигается максимальное качество прогноза смещается в сторону бо́льших значений при увеличении радиуса окна. Это следствие алгоритма формирования MNA дескрипторов: дескрипторы меньшего порядка входят в состав дескрипторов бо́льшего порядка и более подробно описывают структуру больших соединений. При всех значениях радиуса окна модели показывают хорошее качество предсказание. Все кривые имеют стандартный вид: увеличение качества до некоторого оптимума, после которого наблюдается спад.

По данным внутренней валидации оптимальный радиус окна равен трем, а оптимальный уровень MNA дескрипторов равен девяти. В отличие данных in vitro, данные in vivo не ограничивают нас в использовании окна большего радиуса, однако при радиусе окна, равным четырем наблюдается снижение качества модели на всем диапазоне уровней MNA дескрипторов. Данное снижение может быть связано с алгоритмом получения отрицательных примеров. В качестве отрицательного примера использовался случайный остаток, входящий в состав эпитопа, но не входящий в окрестность положительного примера. Так как эпитопы, в большинстве своем небольшого размера – длиной восемь или девять аминокислотных остатков – то при использовании радиуса окна в четыре остатка для выбора остаются три или четыре аминокислотных остатка, и, следовательно, окрестности положительного и отрицательного примера будут пересекаться. Пересечение окрестностей положительного и отрицательного примеров приводит к искажению оценок вероятности связи с конкретной активностью и росту ложных предсказаний.



**Рисунок 9.** Зависимость метрики внутренней оценки качества моделей на данных in vivo от уровня MNA дескрипторов при различных радиусах окна для различных разбиений исходного набора данных.

Внешняя оценка для моделей, построенных на данных in vivo представлена на рисунке 10. На малых уровнях MNA дескрипторов модели с радиусом окна, равного двум, показывают себя лучшими, однако с ростом уровня дескрипторов после некоторого оптимума не происходит рост качества, так как один дескриптор высокого уровня может описывать пептид целиком. Таким образом, модель запоминает не закономерности, определяющие сайт расщепления, а конкретные окрестности, которые ассоциированы с сайтом, иными словами модели переобучаются. Поэтому после шестого уровня дескрипторов качество моделей с радиусом окна в два, становится хуже, чем качество моделей с радиусом окна в три. Однако, после оптимального уровня, модели по тем же причинам начинают переобучаться. Качество моделей с радиусом окна в четыре плавно растет до 9 уровня дескрипторов. Возможно, их качество могло бы превзойти качество моделей с меньшими радиусами окон, но это не наблюдается.



**Рисунок 10**. Зависимость внешней метрики качества моделей от уровня MNA дескрипторов.

Внешняя оценка более объективна, чем внутренняя, так как сделана на независимом от обучающей выборки наборе данных. Поэтому при выборе гиперпараметров приоритет отдавался результатам, полученных при пятикратной кросс-валидации. Выбранные гиперпараметры представлены в таблице 3.

**Таблица 3.** Выбранные значения гиперпараметров для моделей, построенных на разных наборах данных для прогноза сайтов расщепления белков протеасомой.

| **Набор данных** | **Оптимальный радиус окна** | **Оптимальный уровень MNA дескрипторов** |
| --- | --- | --- |
| In vitro, 20S конститутивная протеасома | 2 | 7 |
| In vitro, 20S, иммунопротеасома | 2 | 7 |
| In vivo | 3 | 9 |

В таблице 4 представлены рассчитанные статистические метрики по результатам пятикратной кросс-валидации для моделей с оптимальными гиперпараметрами.

**Таблица 4**. Статистические характеристики моделей, построенных на разных наборах данных для прогноза сайтов расщепления белков протеасомой с оптимальными гиперпараметрами

| **Метрика** | **In vitro, 20S конститутивная протеасома** | **In vitro, 20S, иммунопротеасома** | **In vivo** |
| --- | --- | --- | --- |
| Порог | -0,08 | 0,48 | -0,03 |
| AUROC | 0,68 | 0,63 | 0,87 |
| Чувствительность | 0,69 | 0,43 | 0,81 |
| Специфичность | 0,58 | 0,81 | 0,78 |
| Точность | 0,63 | 0,68 | 0,80 |
| Сбалансированная точность | 0,63 | 0,62 | 0,80 |
| Средняя точность | 0,59 | 0,50 | 0,86 |
| Коэффициент корреляции Мэтью | 0,27 | 0,25 | 0,60 |

По данным таблицы 5 можно сделать следующие промежуточные выводы:

1. Модели, построенные по данным in vivo, превосходят по всем характеристикам модели, построенные по данным in vitro;
2. Модели конститутивной протеасомы лучше моделей иммунопротеасомы;
3. Модели, построенные на данных in vitro не применимы для практического использования;
4. Модели, построенные на данных in vivo, имеют хорошее качество.

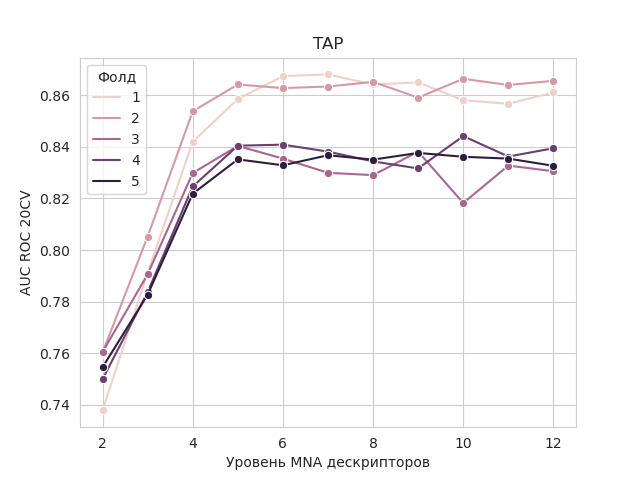
Для дальнейшей работы и построения прототипа веб-сервиса была взята модель, построенная на данных in vivo с радиусом окна, равного трём и уровнем MNA дескрипторов, равного девяти.

## 3.2 Моделирование TAP

Набор данных, использованный для прогноза характера активности пептидов с TAP транспортером, содержит 613 пептидов. Количество сильноактивных пептидов - 468, слабоактивных или неактивных - 145. Для моделей предсказания характера активности к TAP транспортеру, также как и для моделей предсказания связывания с аллелями HLA необходимо оптимизировать всего лишь один гиперпараметр – уровень MNA дескрипторов. Результаты внутренней оценки представлены на рисунке 11. Стоит отметить необычное поведение кривых на рисунке 11. Пять кривых после четвертого уровня MNA дескрипторов, разделяются на две группы с выходом на плато у значения AUC 0,84 и 0,86. Кривые имеют гиперболический вид с выходом на плато после пятого уровня дескрипторов для одной группы кривых и после шестого уровня дескрипторов для другой группы кривых. Исходя из этого можно сделать два вывода:

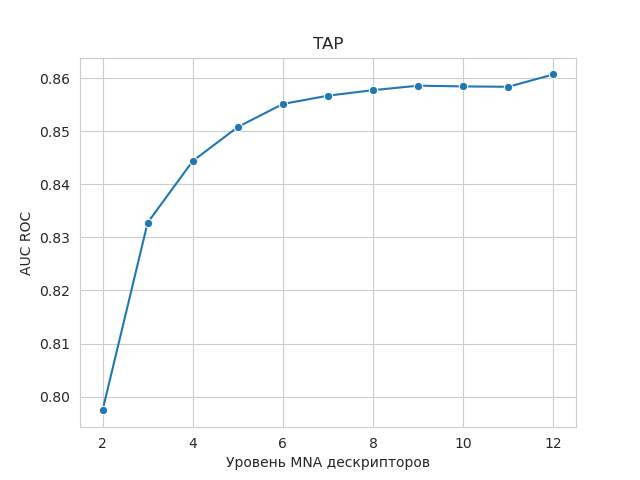
1. Разбиения несбалансированы. Можно разбить их на две группы, в одной из которых присутствуют пептиды, содержащие дескрипторы, которые сильно влияют на качество предсказания.
2. Использование уровня MNA дескрипторов больше шести нецелесообразно, так как модели с использованием более высокого порядка дескрипторов переобучаются.

Характерная длина пептидов в обучающих и тестовых наборах равна девяти. Механизм переобучения заключается в запоминании моделью ассоциации крупных молекулярных фрагментов с характером активности, а не выявлении физико-химических особенностей молекулы, ответственных за характер активности.



**Рисунок 11.** Зависимость метрики внутренней оценки качества моделей TAP транспортера от уровня MNA для различных разбиений исходного набора данных.

Результаты внешней оценки представлены на рисунке 12. Кривая имеет гиперболический вид, как и при внутренней оценке качества. Как было сказано выше, это может свидетельствовать о переобучении, поэтому в качестве оптимального уровня дескриптора была взята первая точка на плато, а не максимум. После шестого уровня рост качества моделей не происходит или происходит незначительно. Это значения уровня было выбрано как оптимальное. Статистические характеристики модели, построенной с использованием шестого уровня MNA дескрипторов, представлены в таблице 5.



**Рисунок 12.** Зависимость метрики внешней оценки качества моделей TAP транспортера от уровня MNA дескрипторов.

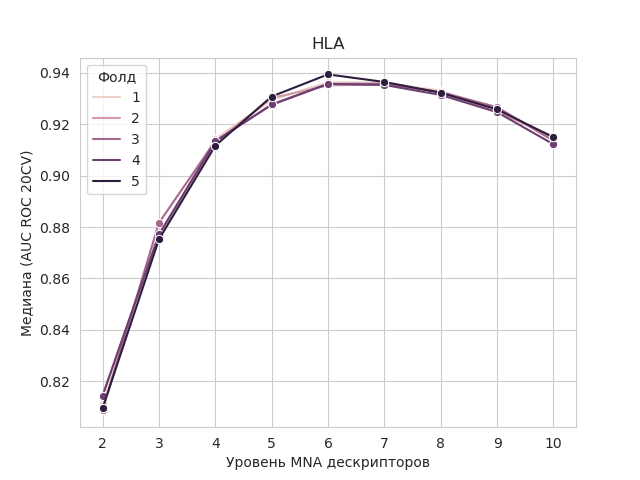
**Таблица 5**. Статистические характеристики моделей TAP транспортера с оптимальным уровнем MNA дескрипторов

| **Метрика** | **Значение** |
| --- | --- |
| Оптимальный порог | 0,02 |
| AUROC | 0,82 |
| Чувствительность | 0,72 |
| Специфичность | 0,83 |
| Точность | 0,72 |
| Сбалансированная точность | 0,77 |
| Средняя точность | 0,98 |
| Коэффициент корреляции Мэтью | 0,31 |

Несмотря на хорошие статистические характеристики моделей. Можно ожидать большую ошибку генерализации моделей, так как они обучались на ограниченном по всем параметрам наборе данных: количестве пептидов, длине пептидов, аминокислотному составу, размаху варьирования IC50. Для дальнейшей работы и построения прототипа веб-сервиса была взята модель, построенная с использованием уровня MNA дескрипторов, равного шести.

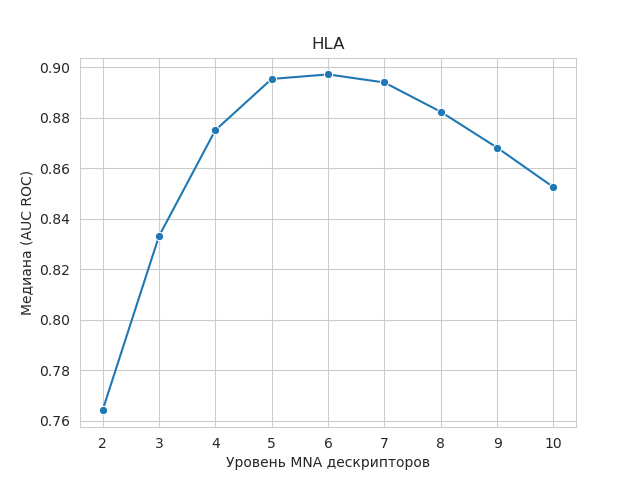
## 3.3 Моделирование взаимодействия с HLA I

Готовый набор данных, использованный для прогноза характера активности пептидов с HLA I, содержит 416084 пептида и 193 аллеля. Для интегральной оценки качества моделей и подбора оптимального уровня MNA дескрипторов, единственного гиперапараметра, использовалась медиана от ряда значений AUROC для каждой аллели. Медиана была выбрана как легко интерпретируемый показатель – связывание половины аллелей предсказывается лучше, чем некоторое число. Помимо этого, медиана более устойчива к выбросам, чем среднее арифметическое. Аллели, для которых метрика внутреннего качества получилась меньше 0,7 были отфильтрованы. Результаты внутренней оценки представлены на рисунке 13. Кривые имеют типичный вид, максимум достигается при использовании шестого уровня дескрипторов, что согласуется с результатами моделирования TAP транспортера, так как характерная длина пептидов в обучающей и тестовой выборке также равна девяти. Кривые имеют одинаковый вид и расположены близко друг к другу на всем интервале варьирования гиперпараметра. Это говорит о хорошей сбалансированности разбиений данных. Не все аллели проходили внутренний контроль качества. Итоговое количество предсказываемых аллелей по результатам внутренней оценки качества варьирует от 176 до 179. Для шестого уровня количество предсказываемых аллелей равно 176.



**Рисунок 13**. Зависимость медианы метрики внутренней оценки качества моделей прогноза связывания с HLA I от уровня MNA для различных разбиений исходного набора данных.

Результаты внешней оценки представлены на рисунке 14. Кривая имеет такой же вид, как и кривые внутренней оценки качества, только несколько смещена в сторону меньших значений метрики качества моделей. Оптимальное значение уровня MNA дескрипторов также равно шести. Дальнейшее уменьшении метрик на высоких уровнях может быть следствием переобучения моделей.Аллели, для которых метрика внешнего качества получилась меньше 0,7 были также исключены из итоговой модели. Количество пептидов и значения статистических характеристик для каждой прогнозируемой аллели приведено в Приложении 2.



**Рисунок 14.** Зависимость медианы метрики внешней оценки качества моделей прогноза связывания с HLA I от уровня MNA дескрипторов.

Медианные статистические характеристики модели, построенной с использованием шестого уровня MNA дескрипторов представлена в таблице 6.

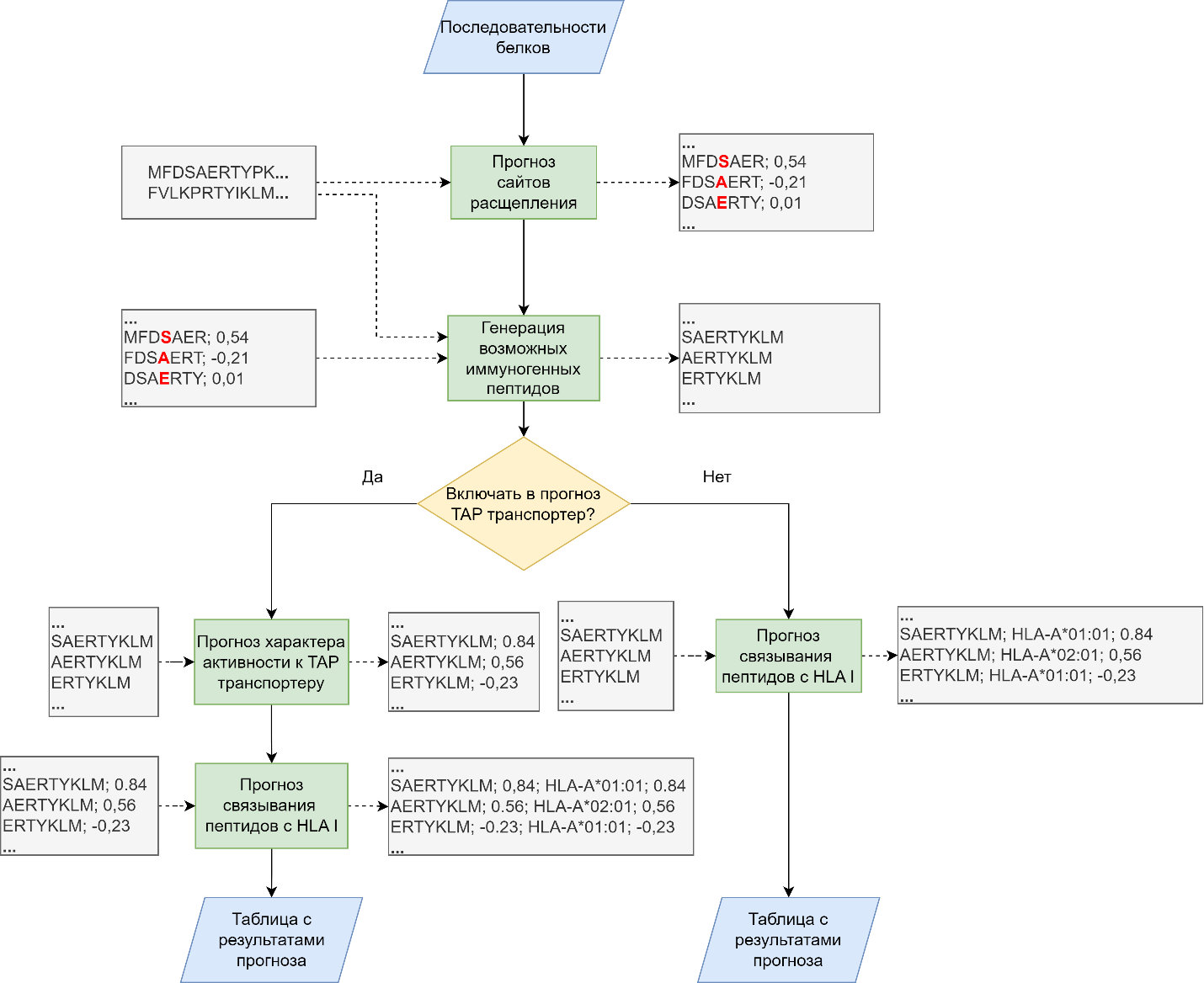
**Таблица 6.** Статистические характеристики моделей прогноза связывания с HLA I с оптимальным уровнем MNA дескрипторов. Все метрики, кроме количества прогнозируемых аллелей представлены как медиана значений от всех классов (аллелей).

| **Метрика** | **Значение** |
| --- | --- |
| Количество прогнозируемых аллелей | 171 |
| Оптимальный порог | 0,04 |
| AUROC | 0,90 |
| Чувствительность | 0,85 |
| Специфичность | 0,90 |
| Точность | 0,90 |
| Сбалансированная точность | 0,86 |
| Средняя точность | 0,07 |
| Коэффициент корреляции Мэтью | 0,14 |

Для дальнейшей работы и построения прототипа веб-сервиса была взята модель, построенная с использованием уровня MNA дескрипторов, равного шести.

## 3.4. Веб-сервис

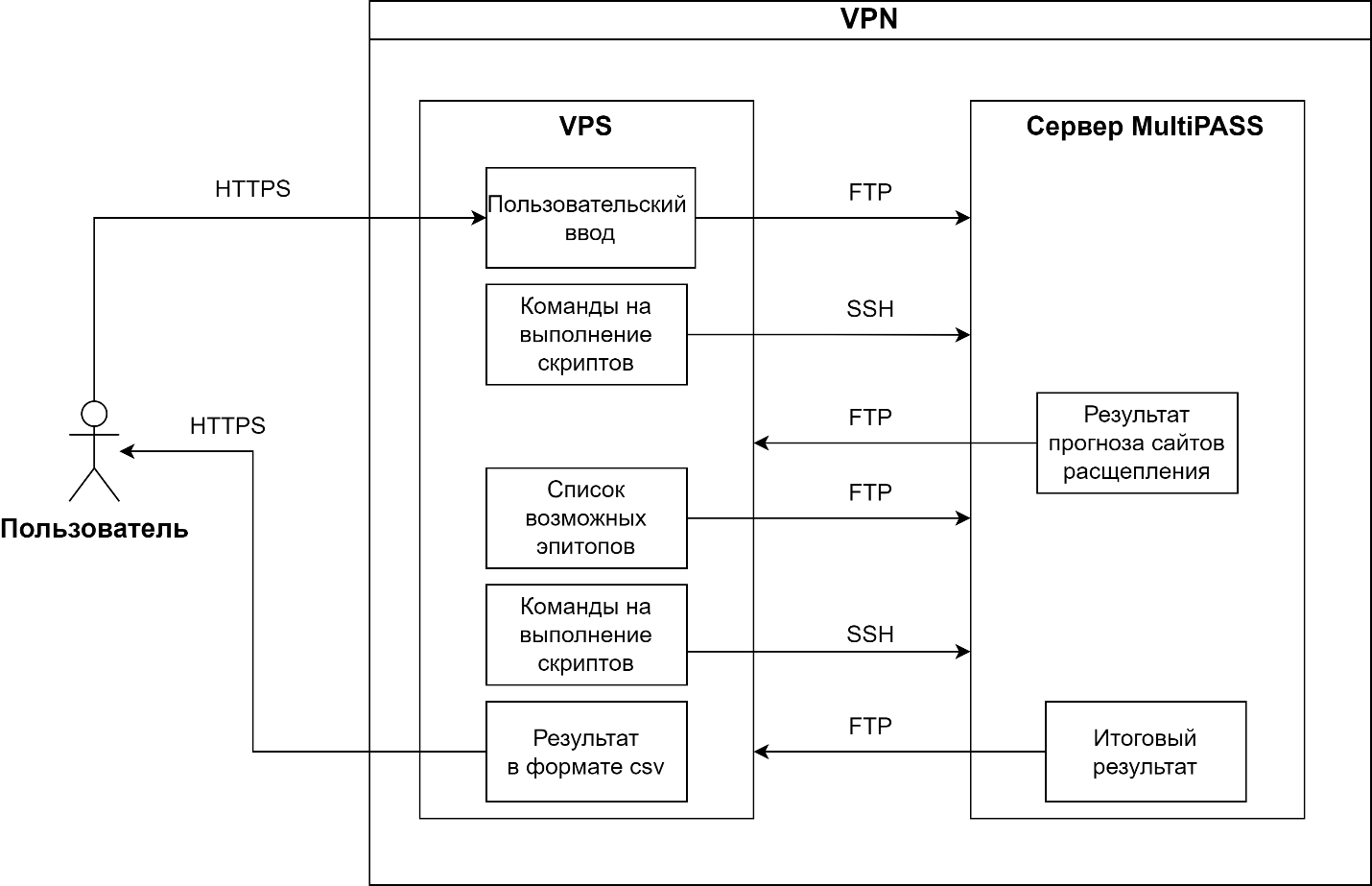
Для построение единой системы для предсказания эпитопов были взяты модели предсказания расщепления белков протеасомой, обученной на данных in vivo, модель прогноза характера активности пептидов по отношению к TAP транспортеру, модель прогноза связывания пептидов с аллелями HLA. Ввиду того, что модель TAP транспортера из-за особенностей её построения и ожидаемой большой ошибки генерализации не может служить полноценным фильтром для предыдущих этапов работы, то её включение в прогноз сделано опциональным. Программе было присвоено название ANGEL (ANtigen GEneration). Принципиальная схема её работы представлена на рисунке 15.



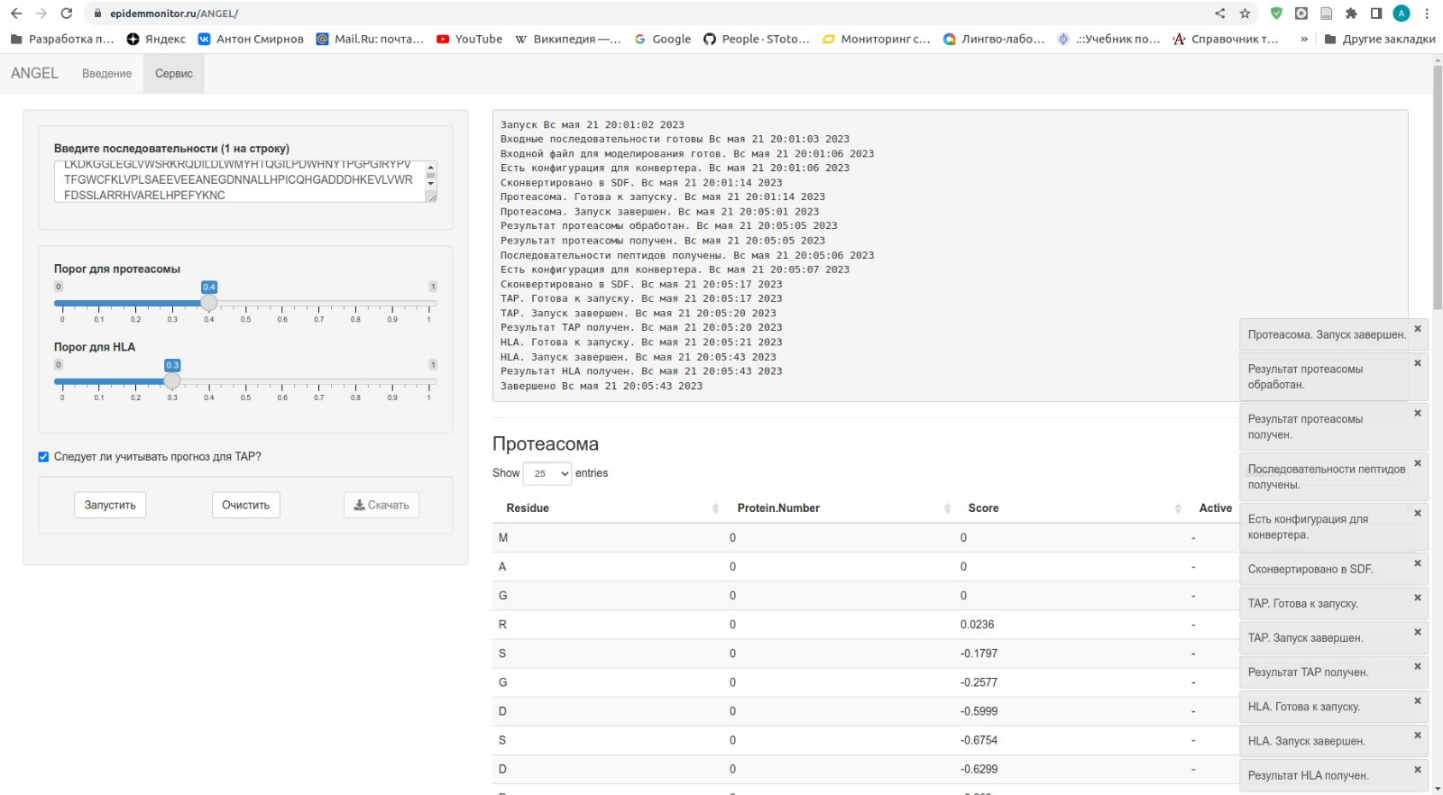
**Рисунок 15.** Принципиальная схема работы веб-сервиса с примерами входных и выходных данных.

Для первого этапа входные последовательности белков разбиваются на подпоследовательности скользящим окном радиусом в 3 аминокислотных остатка. Эти подпоследовательности конвертируются в SDF формат. Создается конфигурационный файл, в котором прописывается путь к SDF файлу, путь к файлу модели, и путь к выходному файлу. Следующим шагом запускается консольная версия программы MultiPASS, которая получает на вход этот конфигурационный файл. Используя полученный файл с прогнозом и исходный файл с входными последовательностями, ANGEL получает прогноз для каждого остатка, присваивая крайним 3 остаткам с N и C-концов последовательностей значение ноль. Затем исходные последовательности делятся на подпоследовательности таким образом, чтобы остаток, для которого прогноз выше или равен некоторому порогу, заданным пользователем, оказался С-концом. Получившиеся подпоследовательности длиной больше семи аминокислот итеративно укорачиваются с N-конца с шагом в один остаток до минимальной длины в восемь остатков и фильтруются все последовательности длиной больше 13 остатков. Таким образом, моделируется работа других протеаз, которые могут воздействовать на вероятный эпитоп. Полученное множество последовательностей конвертируется в SDF формат. В зависимости от выбора пользователя генерируется конфигурационный файлы для запуска прогноза характера активности к TAP транспортеру и HLA I, тогда в выходном файле будет объединение двух прогнозов, или только для запуска прогноза активности к HLA I.

Из-за технических ограничений программ MultiPASS и Shiny Server веб-сервис работает через взаимодействие двух машин. Поток информации между узлами веб-сервиса и пользователем изображен на рисунке 16. Веб-сервис развернут на удаленном VPS, свободно доступен по адресу <https://epidemmonitor.ru/ANGEL/>. Он способен работать в многопользовательском режиме, но не рассчитан на большое количество пользователей из-за ограниченных вычислительных возможностей. Пример работы изображен на рисунке 17. Для пользователя предоставляется поле для ввода белковых последовательностей. Одна последовательность должна быть записана в одну строку. Прототип осуществляет проверку ввода на наличие недопустимых символов и минимальную длину последовательности, равной восьми. Ползунки необходимы для выбора порогов активности, галочка для включения прогноза характера активности к TAP транспортеру. На выход программа выдает две таблицы: прогноз сайтов расщепления белков протеасомой и прогноз, какие вероятные эпитопы с какими аллелями HLA будет связываться. Эти таблицы доступны для скачивания пользователем в формате csv.



**Рисунок 16.** Потоки информации между узлами веб-сервиса и пользователем. Над стрелками подписаны протоколы, по которым осуществляется обмен информацией



**Рисунок 17.** Пример работы прототипа

## 3.5. Обсуждение

Полученные результаты имеют неоднозначный характер. С одной стороны, показана возможность моделирования процессов антигенного процессинга, используя информацию о химической структуре молекул белков на достаточно высоком уровне. С другой стороны, существуют некоторые проблемы. Первая проблема – тяжело найти независимую выборку для непредвзятой валидации моделей и сравнения с существующими аналогами. Для моделей HLA существуют два недавних исследования, в которых проводится сравнение существующих аналогов [99, 102]. Однако только авторы[102] предоставляют итоговый датасет, но он почти полностью пересекается с обучающей выборкой, использованной в данной работе. В целом, описания подготовки датасетов и получившиеся результаты вызывают подозрения в качестве этих работ. Для моделей TAP существует крайне малое количество данных. Кроме использованного в этой работе датасета существует БД MHC BN 4.0 [78]. Извлечение данных из этой БД составило отдельную техническую задачу. Однако её содержимое тоже сильно пересекается также сильно пересекается с использованной обучающей выборкой. На полученном наборе данных из 61 одного пептида модель для прогноза характера связывания пептидов TAP транспортером показала слабый перфоманс (AUROC = 0,60). Это также одна из причин, почему её включение в работу сервиса отдано на усмотрение пользователя. На найденных независимых данных по прогнозу сайтов расщепления белков протеасомой также наблюдалось существенное падение перфоманса (AUROC для модели in vivo = 0,72). Возможно, это связано с двумя причинами. Первая причина - используемые в обучающей выборке пептиды слишком малого размера и необходимо брать дескрипторы более высокого порядка. Вторая причина – обучающий датасет был искусственно сбалансирован: алгоритм его формирования предполагает, что на один положительный пример приходится один отрицательный пример, хотя в реальности соотношение активных и неактивных остатков не один к одному, а гораздо больше, потому что из протеасомы могут выходить фрагменты длиной и в 5, и в 10, и в 15, и с ещё большим количеством аминокислот. Здесь мы подходим к второй проблеме – технические ограничения программы MultiPASS. У неё экспериментально выявлено ограничение на количество дескрипторов равное 16777214, что не позволяет использовать дескрипторы высокого порядка. В самой программе задан максимальный уровень, равный 15. Программа для прогноза сайтов расщепления, показавшая наилучший перфоманс – pepsickle – использует скользящие окна для формирования выборки радиусом в 8 и 21 остаток [100]. Экспериментально проверено, что чем длиннее последовательность, тем бо́льший уровень MNA дескрипторов необходимо использовать. Также определенные сложности создало то, что она работает в однопоточном режиме и только под операционной системой Windows. Из-за отсутствия кроссплатформенности веб-сервис имеет сложную архитектуру. Это можно решить, самостоятельно реализовав алгоритм MultiPASS с использованием современных подходов и библиотек, но это выходит за рамки данной работы.

Третья проблема – отсутствие адекватного количества отрицательных примеров для предсказания связывания с HLA. Адекватное количество есть для данных in vitro, где количество отрицательных примеров несколько большее, чем положительных. Однако данных in vivo, где присутствуют практически только положительные примеры на порядок больше. Перфоманс моделей, обученных на данных in vitro несколько меньше (Средний AUROC = 0,89), чем у моделей, обученных на данных in vivo (Средний AUROC = 0,91). Модели, обученные на смешанных данных, показывают ещё лучшие результаты (Средний AUROC = 0,93). Включение в обучающую выборку снижало качество моделей. MultiPASS считает отрицательными примерами все примеры, которые не относят к обучаемому классу, что является особенностью программы, однако такой подход не верен с биологической точки зрения: один пептид может связываться с разными вариантами HLA. Причиной низкой средней точности (average precision) предсказания является то, для каждого пептида неизвестен истинный спектр связываемых аллелей, а только тот, на который его протестировали, поэтому рекомендуем не ориентироваться на эту метрику.

Обращает на себя внимание плохое качество моделей протеасомы, обученных на in vitro данных. Вероятно, это связано с неоднозначностью присутствующая в данных: слишком много активных сайтов, расположенных близко друг к другу, из-за чего модели могут ошибаться в том, какое окружение относится к активному сайту, а какое – к неактивному.

# Глава 4. Заключение

Разработка методов предсказания Т-клеточных эпитопов является актуальной проблемой иммуноинформатики, так как без них задача создания новых вакцин, иммунотерапевтических и других белковых препаратов становится крайне тяжелой задачей. Последняя пандемия показала, что для человечества необходимо иметь эффективные программы, ускоряющие разработку вакцин, программы, которые рассчитывают, как возбудитель взаимодействует с инфицируемым организмом. Важным элементом этого взаимодействия является презентация антигенов на MHC. Существующие методы, несмотря на превосходные значения AUC, не дают ожидаемого от них результата. Необходимость разработки новых методов обосновывается внешнеполитической ситуацией. Например, начиная с мая 2023 года, сайт IEDB не доступен на территории РФ без технических средств обхода блокировок, а этот сайт является агрегатором классических методов, используемых для этой задачи и основополагающим ресурсом для иммунноинформатических исследований.

Задача предсказания Т-клеточных эпитопов является лишь одной стороной медали. Кроме этого, необходимо уметь предсказывать какие Т-клеточные рецепторы будут реагировать на презентируемый эпитоп. Это отдельная сложная задача, для которой также показано, что использование моделей «структура-активность» на MNA дескрипторах могут показывать хороший результат [116]. Однако все сервисы работают на предсказания активности к ограниченному количеству эпитопов, которые есть в обучающей выборке (проблема «unseen epitopes»). В этой области ведется активная разработка способов преодоления этой проблемы. Современная версия MultiPASS, к сожалению, не способна решить ни проблему «unseen epitopes», ни проблему предсказания активности к всем известным аллелям MHC (проблема «pan-allelic prediction»). Также необходимо уметь предсказывать направление иммунного ответа, скорость и динамику его развития. Отдельной сложной задачей является предсказания B-клеточных эпитопов. Для их предсказания необходимо знать трехмерную структуру антигена и сайты связывания антитела. Такие эксперименты являются более дорогими и трудоемкими, и по ним не накоплено большого количества данных. Для разработки белковых препаратов, вакцин является важным предсказание аллергенности – ещё одна сложная задача, так как накоплено мало данных, связывающих аллергическую реакцию с конкретным белком, её вызвавшую, и необходимо уметь хорошо предсказывать B-клеточные эпитопы.

Несмотря на ряд проблем, построенные модели показали хорошее качество. В дальнейшем планируется работа по их преодолению. Кроме того, необходимо построить модели антигенного процессинга для MHC второго класса и интегрировать модели антигенного процессинга с моделями предсказания активности Т-клеточных рецепторов и инструментами анализа репертуара ТКР. В дальнейшем будет сделано прямое сравнение современных существующих методов с разрабатываемым.

# Глава 5. Выводы

По итогам проделанной работы можно сделать следующие выводы:

1. Были подготовлены выборки для обучения и валидации моделей прогноза сайтов расщепления белков протеасомой: по данным in vitro для конститутивной протеасомы (88 белков), для иммуннопротеасомы (24 белка), по данным in vivo (56 187 белков); модели прогноза активности белковых фрагментов по отношению к ТАР транспортеру (613 пептидов); модели прогноза связывания белковых фрагментов с конкретными аллелями MHC I (416084 пептида для 193 аллелей).
2. Были построены и валидированы модели «структура-активность» для прогноза сайтов расщепления белков протеасомой: по данным in vitro для конститутивной протеасомы (AUROC = 0,68), для иммуннопротеасомы (AUROC = 0,63), по данным in vivo (AUROC = 0,87); для прогноза активности белковых фрагментов по отношению к ТАР транспортеру (AUROC = 0,82); для прогноза связывания белковых фрагментов с конкретными аллелями MHC I (медиана AUROC = 0,90 для 171 аллели).
3. Построенные модели были объединены в единую систему, конвейер для предсказания эпитопов, презентируемых на MHC первого класса
4. Развернут веб-сервис для свободного доступа к разработанной системе через сеть Интернет.

# Список литературы

1. Ярилин А.А.Иммунология / А. А. Ярилин – Москва: ГЭОТАР-Медиа, 2010.– 752c.

2. Ганковская Л.В.Основы общей иммунологии. Учебно-методическое пособие для студентов медицинских вузов / Л. В. Ганковская, Л. С. Намазова-Баранова, Р. Я. Мешкова – Москва: ПедиатрЪ, 2014.

3. Murugan A. Statistical inference of the generation probability of T-cell receptors from sequence repertoires / Murugan A., Mora T., Walczak A.M., Callan C.G. // Proceedings of the National Academy of Sciences – 2012. – Т. 109 – № 40 – С.16161–16166.

4. Dash R. In silico-based vaccine design against Ebola virus glycoprotein / Dash R., Das R., Junaid M., Akash M.F.C., Islam A., Hosen S.Z. // Advances and Applications in Bioinformatics and Chemistry – 2017. – Т. Volume 10 – С.11–28.

5. Viana Invenção M. da C. Development of synthetic antigen vaccines for COVID-19 / Viana Invenção M. da C., Melo A.R. da S., Macêdo L.S. de, Costa Neves T.S.P. da, Melo C.M.L. de, Cordeiro M.N., Aragão Batista M.V. de, Freitas A.C. de // Human Vaccines & Immunotherapeutics – 2021. – Т. 17 – № 11 – С.3855–3870.

6. Schmidt J. Prediction of neo-epitope immunogenicity reveals TCR recognition determinants and provides insight into immunoediting / Schmidt J., Smith A.R., Magnin M., Racle J., Devlin J.R., Bobisse S., Cesbron J., Bonnet V., Carmona S.J., Huber F., Ciriello G., Speiser D.E., Bassani-Sternberg M., Coukos G., Baker B.M., Harari A., Gfeller D. // Cell Reports Medicine – 2021. – Т. 2 – № 2 – С.100194.

7. consortium M. sequencing Complete sequence and gene map of a human major histocompatibility complex / consortium M. sequencing // Nature – 1999. – Т. 401 – № 6756 – С.921–923.

8. Robinson J. IPD-IMGT/HLA Database / Robinson J., Barker D.J., Georgiou X., Cooper M.A., Flicek P., Marsh S.G.E. // Nucleic Acids Research – 2019.

9. IPD-IMGT/HLA Database Statistics [Электронный ресурс]. URL: https://www.ebi.ac.uk/ipd/imgt/hla/about/statistics/ (дата обращения: 27.02.2023).

10. Cao K. Analysis of the frequencies of HLA-A, B, and C alleles and haplotypes in the five major ethnic groups of the United States reveals high levels of diversity in these loci and contrasting distribution patterns in these populations / Cao K., Hollenbach J., Shi X., Shi W., Chopek M., Fernández-Viña M.A. // Human Immunology – 2001. – Т. 62 – № 9 – С.1009–1030.

11. Autoimmunity: From Bench to Bedside / / под ред. J.-M. Anaya, Y. Shoenfeld, A. Rojas-Villarraga, R.A. Levy, R. Cervera. – – Bogota (Colombia): El Rosario University Press, 2013.

12. Мерфи К.Иммунобиология по Джанвею / К. Мерфи, К. Уивер – Москва: Логосфера, 2020. Вып. 9– 1184c.

13. Lanier L.L. NK CELL RECOGNITION / Lanier L.L. // Annual Review of Immunology – 2005. – Т. 23 – № 1 – С.225–274.

14. Alfonso C. Nonclassical MHC Class II Molecules / Alfonso C., Karlsson L. // Annual Review of Immunology – 2000. – Т. 18 – № 1 – С.113–142.

15. Тибирькова Е.В. РОЛЬ ИММУНОЛОГИЧЕСКОЙ ТОЛЕРАНТНОСТИ В ПОДДЕРЖАНИИ ГОМЕОСТАЗА / Тибирькова Е.В., Белан Э.Б., Желтова А.А., Садчикова Т.Л.

16. Deakin J.E. Evolution and comparative analysis of the MHC Class III inflammatory region / Deakin J.E., Papenfuss A.T., Belov K., Cross J.G., Coggill P., Palmer S., Sims S., Speed T.P., Beck S., Graves J.A.M. // BMC Genomics – 2006. – Т. 7 – № 1 – С.281.

17. Bjorkman P.J. The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens / Bjorkman P.J., Saper M.A., Samraoui B., Bennett W.S., Strominger J.L., Wiley D.C. // Nature – 1987. – Т. 329 – № 6139 – С.512–518.

18. Meusser B. ERAD: the long road to destruction / Meusser B., Hirsch C., Jarosch E., Sommer T. // Nature Cell Biology – 2005. – Т. 7 – № 8 – С.766–772.

19. Trolle T. The Length Distribution of Class I–Restricted T Cell Epitopes Is Determined by Both Peptide Supply and MHC Allele–Specific Binding Preference / Trolle T., McMurtrey C.P., Sidney J., Bardet W., Osborn S.C., Kaever T., Sette A., Hildebrand W.H., Nielsen M., Peters B. // The Journal of Immunology – 2016. – Т. 196 – № 4 – С.1480–1487.

20. Bjorkman P.J. Structures of two classes of MHC molecules elucidated: crucial differences and similarities / Bjorkman P.J., Burmeister W.P. // Current Opinion in Structural Biology – 1994. – Т. 4 – № 6 – С.852–856.

21. White K.D. HLA and the Pharmacogenomics of Drug Hypersensitivity Elsevier, 2014. – 437–465с.

22. Matsumura M. Emerging Principles for the Recognition of Peptide Antigens by MHC Class I Molecules / Matsumura M., Fremont D.H., Peterson P.A., Wilson lan A. // Science – 1992. – Т. 257 – № 5072 – С.927–934.

23. Bjorkman P.J. Structures of two classes of MHC molecules elucidated: crucial differences and similarities / Bjorkman P.J., Burmeister W.P. // Current Opinion in Structural Biology – 1994. – Т. 4 – № 6 – С.852–856.

24. Hewitt E.W. The MHC class I antigen presentation pathway: strategies for viral immune evasion / Hewitt E.W. // Immunology – 2003. – Т. 110 – № 2 – С.163–169.

25. Rutigliano H.M. Trophoblast Major Histocompatibility Complex Class I Expression Is Associated with Immune-Mediated Rejection of Bovine Fetuses Produced by Cloning / Rutigliano H.M., Thomas A.J., Wilhelm A., Sessions B.R., Hicks B.A., Schlafer D.H., White K.L., Davies C.J. // Biology of Reproduction – 2016. – Т. 95 – № 2 – С.39–39.

26. Ting J.P.-Y. Genetic Control of MHC Class II Expression / Ting J.P.-Y., Trowsdale J. // Cell – 2002. – Т. 109 – № 2 – С.S21–S33.

27. Rock K.L. Present Yourself! By MHC Class I and MHC Class II Molecules / Rock K.L., Reits E., Neefjes J. // Trends in Immunology – 2016. – Т. 37 – № 11 – С.724–737.

28. Bevan M.J. Cross-priming / Bevan M.J. // Nature Immunology – 2006. – Т. 7 – № 4 – С.363–365.

29. Pishesha N. A guide to antigen processing and presentation / Pishesha N., Harmand T.J., Ploegh H.L. // Nature Reviews Immunology – 2022. – Т. 22 – № 12 – С.751–764.

30. Gong B. The Ubiquitin-Proteasome System: Potential Therapeutic Targets for Alzheimer’s Disease and Spinal Cord Injury / Gong B., Radulovic M., Figueiredo-Pereira M.E., Cardozo C. // Frontiers in Molecular Neuroscience – 2016. – Т. 9.

31. Tanaka K. The proteasome: Overview of structure and functions / Tanaka K. // Proceedings of the Japan Academy, Series B – 2009. – Т. 85 – № 1 – С.12–36.

32. Kloetzel P.-M. Antigen processing by the proteasome / Kloetzel P.-M. // Nature Reviews Molecular Cell Biology – 2001. – Т. 2 – № 3 – С.179–188.

33. Murata S. The immunoproteasome and thymoproteasome: functions, evolution and human disease / Murata S., Takahama Y., Kasahara M., Tanaka K. // Nature Immunology – 2018. – Т. 19 – № 9 – С.923–931.

34. Murata S. Regulation of CD8 + T Cell Development by Thymus-Specific Proteasomes / Murata S., Sasaki K., Kishimoto T., Niwa S., Hayashi H., Takahama Y., Tanaka K. // Science – 2007. – Т. 316 – № 5829 – С.1349–1353.

35. Harris J.L. Substrate specificity of the human proteasome / Harris J.L., Alper P.B., Li J., Rechsteiner M., Backes B.J. – С.11.

36. Takahama Y. β5t-containing thymoproteasome: specific expression in thymic cortical epithelial cells and role in positive selection of CD8+ T cells / Takahama Y., Takada K., Murata S., Tanaka K. // Current Opinion in Immunology – 2012. – Т. 24 – № 1 – С.92–98.

37. Kniepert A. The unique functions of tissue-specific proteasomes / Kniepert A., Groettrup M. // Trends in Biochemical Sciences – 2014. – Т. 39 – № 1 – С.17–24.

38. Mediani L. Defective ribosomal products challenge nuclear function by impairing nuclear condensate dynamics and immobilizing ubiquitin / Mediani L., Guillén‐Boixet J., Vinet J., Franzmann T.M., Bigi I., Mateju D., Carrà A.D., Morelli F.F., Tiago T., Poser I., Alberti S., Carra S. // The EMBO Journal – 2019. – Т. 38 – № 15.

39. Dolan B.P. Defective Ribosomal Products Are the Major Source of Antigenic Peptides Endogenously Generated from Influenza A Virus Neuraminidase / Dolan B.P., Li L., Takeda K., Bennink J.R., Yewdell J.W. // The Journal of Immunology – 2010. – Т. 184 – № 3 – С.1419–1424.

40. Suzuki T. The cytoplasmic peptide:N-glycanase (NGLY1) — Structure, expression and cellular functions / Suzuki T., Huang C., Fujihira H. // Gene – 2016. – Т. 577 – № 1 – С.1–7.

41. Kasteren S.I. van Chemical biology of antigen presentation by MHC molecules / Kasteren S.I. van, Overkleeft H., Ovaa H., Neefjes J. // Current Opinion in Immunology – 2014. – Т. 26 – С.21–31.

42. Hanada K. Immune recognition of a human renal cancer antigen through post-translational protein splicing / Hanada K., Yewdell J.W., Yang J.C. // Nature – 2004. – Т. 427 – № 6971 – С.252–256.

43. Admon A. Are There Indeed Spliced Peptides in the Immunopeptidome? / Admon A. // Molecular & Cellular Proteomics – 2021. – Т. 20 – С.100099.

44. Vigneron N. Peptide splicing by the proteasome / Vigneron N., Ferrari V., Stroobant V., Abi Habib J., Van den Eynde B.J. // Journal of Biological Chemistry – 2017. – Т. 292 – № 51 – С.21170–21179.

45. Mansurkhodzhaev A. Proteasome-Generated cis-Spliced Peptides and Their Potential Role in CD8+ T Cell Tolerance / Mansurkhodzhaev A., Barbosa C.R.R., Mishto M., Liepe J. // Frontiers in Immunology – 2021. – Т. 12 – С.614276.

46. Mishto M. An in silico—in vitro Pipeline Identifying an HLA-A\*02:01+ KRAS G12V+ Spliced Epitope Candidate for a Broad Tumor-Immune Response in Cancer Patients / Mishto M., Mansurkhodzhaev A., Ying G., Bitra A., Cordfunke R.A., Henze S., Paul D., Sidney J., Urlaub H., Neefjes J., Sette A., Zajonc D.M., Liepe J. // Frontiers in Immunology – 2019. – Т. 10 – С.2572.

47. Kato K. Characterization of Proteasome-Generated Spliced Peptides Detected by Mass Spectrometry / Kato K., Nakatsugawa M., Tokita S., Hirohashi Y., Kubo T., Tsukahara T., Murata K., Chiba H., Takahashi H., Hirano N., Kanaseki T., Torigoe T. // The Journal of Immunology – 2022. – Т. 208 – № 12 – С.2856–2865.

48. Lichti C.F. Navigating Critical Challenges Associated with Immunopeptidomics-Based Detection of Proteasomal Spliced Peptide Candidates / Lichti C.F., Vigneron N., Clauser K.R., Van den Eynde B.J., Bassani-Sternberg M. // Cancer Immunology Research – 2022. – Т. 10 – № 3 – С.275–284.

49. Liepe J. A large fraction of HLA class I ligands are proteasome-generated spliced peptides / Liepe J., Marino F., Sidney J., Jeko A., Bunting D.E., Sette A., Kloetzel P.M., Stumpf M.P.H., Heck A.J.R., Mishto M. // Science – 2016. – Т. 354 – № 6310 – С.354–358.

50. Vigneron N. Production of spliced peptides by the proteasome / Vigneron N., Stroobant V., Ferrari V., Abi Habib J., Van den Eynde B.J. // Molecular Immunology – 2019. – Т. 113 – С.93–102.

51. Rock K.L. Protein degradation and the generation of MHC class I-presented peptides Elsevier, 2002. – 1–70с.

52. Craiu A. Two distinct proteolytic processes in the generation of a major histocompatibility complex class I-presented peptide / Craiu A., Akopian T., Goldberg A., Rock K.L. // Proceedings of the National Academy of Sciences – 1997. – Т. 94 – № 20 – С.10850–10855.

53. Stoltze L. Two new proteases in the MHC class I processing pathway / Stoltze L., Schirle M., Schwarz G., Schröter C., Thompson M.W., Hersh L.B., Kalbacher H., Stevanovic S., Rammensee H.-G., Schild H. // Nature Immunology – 2000. – Т. 1 – № 5 – С.413–418.

54. Rock K.L. Proteases in MHC Class I Presentation and Cross-Presentation / Rock K.L., Farfán-Arribas D.J., Shen L. // The Journal of Immunology – 2010. – Т. 184 – № 1 – С.9–15.

55. Kunisawa J. The Group II Chaperonin TRiC Protects Proteolytic Intermediates from Degradation in the MHC Class I Antigen Processing Pathway / Kunisawa J., Shastri N. // Molecular Cell – 2003. – Т. 12 – № 3 – С.565–576.

56. Daumke O. Functional asymmetry of the ATP-binding-cassettes of the ABC transporter TAP is determined by intrinsic properties of the nucleotide binding domains: Modular function of TAP domains / Daumke O., Knittler M.R. // European Journal of Biochemistry – 2001. – Т. 268 – № 17 – С.4776–4786.

57. Urlinger S. Intracellular Location, Complex Formation, and Function of the Transporter Associated with Antigen Processing in Yeast / Urlinger S., Kuchler K., Meyer T.H., Uebel S., Tampe R. // European Journal of Biochemistry – 1997. – Т. 245 – № 2 – С.266–272.

58. Blees A. Assembly of the MHC I peptide-loading complex determined by a conserved ionic lock-switch / Blees A., Reichel K., Trowitzsch S., Fisette O., Bock C., Abele R., Hummer G., Schäfer L.V., Tampé R. // Scientific Reports – 2015. – Т. 5 – № 1 – С.17341.

59. Lehnert E. Structure and Dynamics of Antigenic Peptides in Complex with TAP / Lehnert E., Tampé R. // Frontiers in Immunology – 2017. – Т. 8.

60. Abele R. The ABCs of Immunology: Structure and Function of TAP, the Transporter Associated with Antigen Processing / Abele R., Tampé R. // Physiology – 2004. – Т. 19 – № 4 – С.216–224.

61. Grossmann N. Mechanistic determinants of the directionality and energetics of active export by a heterodimeric ABC transporter / Grossmann N., Vakkasoglu A.S., Hulpke S., Abele R., Gaudet R., Tampé R. // Nature Communications – 2014. – Т. 5 – № 1 – С.5419.

62. Herget M. Purification and Reconstitution of the Antigen Transport Complex TAP / Herget M., Kreiβig N., Kolbe C., Schölz C., Tampé R., Abele R. // Journal of Biological Chemistry – 2009. – Т. 284 – № 49 – С.33740–33749.

63. Endert P.M. van A sequential model for peptide binding and transport by the transporters associated with antigen processing / Endert P.M. van, Tampé R., Meyer T.H., Tisch R., Bach J.-F., McDevitt H.O. // Immunity – 1994. – Т. 1 – № 6 – С.491–500.

64. Gorbulev S. Allosteric crosstalk between peptide-binding, transport, and ATP hydrolysis of the ABC transporter TAP / Gorbulev S., Abele R., Tampé R. // Proceedings of the National Academy of Sciences – 2001. – Т. 98 – № 7 – С.3732–3737.

65. Herget M. Conformation of peptides bound to the transporter associated with antigen processing (TAP) / Herget M., Baldauf C., Schölz C., Parcej D., Wiesmüller K.-H., Tampé R., Abele R., Bordignon E. // Proceedings of the National Academy of Sciences – 2011. – Т. 108 – № 4 – С.1349–1354.

66. Androlewicz M.J. Human transporters associated with antigen processing possess a promiscuous peptide-binding site / Androlewicz M.J., Cresswell P. // Immunity – 1994. – Т. 1 – № 1 – С.7–14.

67. Uebel S. Recognition principle of the TAP transporter disclosed by combinatorial peptide libraries / Uebel S., Kraas W., Kienle S., Wiesmüller K.-H., Jung G., Tampé R. // Proceedings of the National Academy of Sciences – 1997. – Т. 94 – № 17 – С.8976–8981.

68. Blees A. Structure of the human MHC-I peptide-loading complex / Blees A., Januliene D., Hofmann T., Koller N., Schmidt C., Trowitzsch S., Moeller A., Tampé R. // Nature – 2017. – Т. 551 – № 7681 – С.525–528.

69. Serwold T. ERAAP customizes peptides for MHC class I molecules in the endoplasmic reticulum / Serwold T., Gonzalez F., Kim J., Jacob R., Shastri N. // Nature – 2002. – Т. 419 – № 6906 – С.480–483.

70. Yewdell J.W. Don’t mess with ERAAP! / Yewdell J.W., Lu X. // Nature Immunology – 2012. – Т. 13 – № 6 – С.526–528.

71. Osborne A.R. PROTEIN TRANSLOCATION BY THE SEC61/SECY CHANNEL / Osborne A.R., Rapoport T.A., Berg B. van den // Annual Review of Cell and Developmental Biology – 2005. – Т. 21 – № 1 – С.529–550.

72. Vita R. The Immune Epitope Database (IEDB): 2018 update / Vita R., Mahajan S., Overton J.A., Dhanda S.K., Martini S., Cantrell J.R., Wheeler D.K., Sette A., Peters B. // Nucleic Acids Research – 2019. – Т. 47 – № D1 – С.D339–D343.

73. Roetschke H.P. InvitroSPI and a large database of proteasome-generated spliced and non-spliced peptides / Roetschke H.P., Rodriguez-Hernandez G., Cormican J.A., Yang X., Lynham S., Mishto M., Liepe J. // Scientific Data – 2023. – Т. 10 – № 1 – С.18.

74. Willimsky G. In vitro proteasome processing of neo-splicetopes does not predict their presentation in vivo / Willimsky G., Beier C., Immisch L., Papafotiou G., Scheuplein V., Goede A., Holzhütter H.-G., Blankenstein T., Kloetzel P.M. // eLife – 2021. – Т. 10 – С.e62019.

75. Toes R.E.M. Discrete Cleavage Motifs of Constitutive and Immunoproteasomes Revealed by Quantitative Analysis of Cleavage Products / Toes R.E.M., Nussbaum A.K., Degermann S., Schirle M., Emmerich N.P.N., Kraft M., Laplace C., Zwinderman A., Dick T.P., Müller J., Schönfisch B., Schmid C., Fehling H.-J., Stevanovic S., Rammensee H.G., Schild H. // Journal of Experimental Medicine – 2001. – Т. 194 – № 1 – С.1–12.

76. Emmerich N.P.N. The Human 26 S and 20 S Proteasomes Generate Overlapping but Different Sets of Peptide Fragments from a Model Protein Substrate / Emmerich N.P.N., Nussbaum A.K., Stevanovic S., Priemer M., Toes R.E.M., Rammensee H.-G., Schild H. // Journal of Biological Chemistry – 2000. – Т. 275 – № 28 – С.21140–21148.

77. Toseland C.P. AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data / Toseland C.P., Clayton D.J., McSparron H., Hemsley S.L., Blythe M.J., Paine K., Doytchinova I.A., Guan P., Hattotuwagama C.K., Flower D.R. // Immunome Research – 2005. – Т. 1 – № 1 – С.4.

78. Lata S. MHCBN 4.0: A database of MHC/TAP binding peptides and T-cell epitopes / Lata S., Bhasin M., Raghava G.P. // BMC Research Notes – 2009. – Т. 2 – № 1 – С.61.

79. Diez-Rivero C.M. Computational analysis and modeling of cleavage by the immunoproteasome and the constitutive proteasome / Diez-Rivero C.M., Lafuente E.M., Reche P.A. // BMC Bioinformatics – 2010. – Т. 11 – № 1 – С.479.

80. Bagaev D.V. VDJdb in 2019: database extension, new analysis infrastructure and a T-cell receptor motif compendium / Bagaev D.V., Vroomans R.M.A., Samir J., Stervbo U., Rius C., Dolton G., Greenshields-Watson A., Attaf M., Egorov E.S., Zvyagin I.V., Babel N., Cole D.K., Godkin A.J., Sewell A.K., Kesmir C., Chudakov D.M., Luciani F., Shugay M. // Nucleic Acids Research – 2020. – Т. 48 – № D1 – С.D1057–D1062.

81. Tickotsky N. McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences / Tickotsky N., Sagiv T., Prilusky J., Shifrut E., Friedman N. // Bioinformatics – 2017. – Т. 33 – № 18 – С.2924–2929.

82. Rammensee H.-G. SYFPEITHI: database for MHC ligands and peptide motifs / Rammensee H.-G., Bachmann J., Emmerich N.P.N., Bachor O.A., Stevanović S. // Immunogenetics – 1999. – Т. 50 – № 3–4 – С.213–219.

83. Reche P.A. EPIMHC: a curated database of MHC-binding peptides for customized computational vaccinology / Reche P.A., Zhang H., Glutting J.-P., Reinherz E.L. // Bioinformatics – 2005. – Т. 21 – № 9 – С.2140–2141.

84. Nussbaum A.K. PAProC: a prediction algorithm for proteasomal cleavages available on the WWW / Nussbaum A.K., Kuttler C., Hadeler K.-P., Rammensee H.-G., Schild H. // Immunogenetics – 2001. – Т. 53 – № 2 – С.87–94.

85. Hattotuwagama C.K. Quantitative online prediction of peptide binding to the major histocompatibility complex / Hattotuwagama C.K., Guan P., Doytchinova I.A., Zygouri C., Flower D.R. // Journal of Molecular Graphics and Modelling – 2004. – Т. 22 – № 3 – С.195–207.

86. Doytchinova I.A. EpiJen: a server for multistep T cell epitope prediction / Doytchinova I.A., Guan P., Flower D.R. // BMC Bioinformatics – 2006. – Т. 7 – № 1 – С.131.

87. Dhanda S.K. IEDB-AR: immune epitope database—analysis resource in 2019 / Dhanda S.K., Mahajan S., Paul S., Yan Z., Kim H., Jespersen M.C., Jurtz V., Andreatta M., Greenbaum J.A., Marcatili P., Sette A., Nielsen M., Peters B. // Nucleic Acids Research – 2019. – Т. 47 – № W1 – С.W502–W506.

88. Bhasin M. Pcleavage: an SVM based method for prediction of constitutive proteasome and immunoproteasome cleavage sites in antigenic sequences / Bhasin M., Raghava G.P.S. // Nucleic Acids Research – 2005. – Т. 33 – № Web Server – С.W202–W207.

89. Nielsen M. The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage / Nielsen M., Lundegaard C., Lund O., Keşmir C. // Immunogenetics – 2005. – Т. 57 – № 1–2 – С.33–41.

90. Gomez-Perosanz M. Identification of CD8+ T cell epitopes through proteasome cleavage site predictions / Gomez-Perosanz M., Ras-Carmona A., Lafuente E.M., Reche P.A. // BMC Bioinformatics – 2020. – Т. 21 – № S17 – С.484.

91. Zhang G. PREDTAP: a system for prediction of peptide binding to the human transporter associated with antigen processing / Zhang G., Petrovsky N., Kwoh C., August J.T., Brusic V. // Immunome Research – 2006. – Т. 2 – № 1 – С.3.

92. Nielsen M. NetMHCpan, a Method for Quantitative Predictions of Peptide Binding to Any HLA-A and -B Locus Protein of Known Sequence / Nielsen M., Lundegaard C., Blicher T., Lamberth K., Harndahl M., Justesen S., Røder G., Peters B., Sette A., Lund O., Buus S. // PLoS ONE – 2007. – Т. 2 – № 8 – С.e796.

93. Phloyphisut P. MHCSeqNet: a deep neural network model for universal MHC binding prediction / Phloyphisut P., Pornputtapong N., Sriswasdi S., Chuangsuwanich E. // BMC Bioinformatics – 2019. – Т. 20 – № 1 – С.270.

94. O’Donnell T.J. MHCflurry 2.0: Improved Pan-Allele Prediction of MHC Class I-Presented Peptides by Incorporating Antigen Processing / O’Donnell T.J., Rubinsteyn A., Laserson U. // Cell Systems – 2020. – Т. 11 – № 1 – С.42- 48.e7.

95. Stranzl T. NetCTLpan: pan-specific MHC class I pathway epitope predictions / Stranzl T., Larsen M.V., Lundegaard C., Nielsen M. // Immunogenetics – 2010. – Т. 62 – № 6 – С.357–368.

96. Wu J. DeepHLApan: A Deep Learning Approach for Neoantigen Prediction Considering Both HLA-Peptide Binding and Immunogenicity / Wu J., Wang W., Zhang J., Zhou B., Zhao W., Su Z., Gu X., Wu J., Zhou Z., Chen S. // Frontiers in Immunology – 2019. – Т. 10 – С.2559.

97. Alvarez B. NNAlign\_MA; MHC Peptidome Deconvolution for Accurate MHC Binding Motif Characterization and Improved T-cell Epitope Predictions / Alvarez B., Reynisson B., Barra C., Buus S., Ternette N., Connelley T., Andreatta M., Nielsen M. // Molecular & Cellular Proteomics – 2019. – Т. 18 – № 12 – С.2459–2477.

98. Paul S. Benchmarking predictions of MHC class I restricted T cell epitopes in a comprehensively studied model system / Paul S., Croft N.P., Purcell A.W., Tscharke D.C., Sette A., Nielsen M., Peters B. // PLOS Computational Biology – 2020. – Т. 16 – № 5 – С.e1007757.

99. Mei S. A comprehensive review and performance evaluation of bioinformatics tools for HLA class I peptide-binding prediction / Mei S., Li F., Leier A., Marquez-Lago T.T., Giam K., Croft N.P., Akutsu T., Smith A.I., Li J., Rossjohn J., Purcell A.W., Song J. // Briefings in Bioinformatics – 2020. – Т. 21 – № 4 – С.1119–1135.

100. Weeder B.R. pepsickle rapidly and accurately predicts proteasomal cleavage sites for improved neoantigen identification / Weeder B.R., Wood M.A., Li E., Nellore A., Thompson R.F. // Bioinformatics – 2021. – Т. 37 – № 21 – С.3723–3733.

101. Daniel S. Relationship Between Peptide Selectivities of Human Transporters Associated with Antigen Processing and HLA Class I Molecules / Daniel S., Brusic V., Caillat-Zucman S., Petrovsky N., Harrison L., Riganelli D., Sinigaglia F., Gallazzi F., Hammer J., Van Endert P.M. // The Journal of Immunology – 1998. – Т. 161 – № 2 – С.617–624.

102. Zhao W. Systematically benchmarking peptide-MHC binding predictors: From synthetic to naturally processed epitopes / Zhao W., Sher X. // PLOS Computational Biology – 2018. – Т. 14 – № 11 – С.e1006457.

103. Paul S. HLA Class I Alleles Are Associated with Peptide-Binding Repertoires of Different Size, Affinity, and Immunogenicity / Paul S., Weiskopf D., Angelo M.A., Sidney J., Peters B., Sette A. // The Journal of Immunology – 2013. – Т. 191 – № 12 – С.5831–5839.

104. Sette A. The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes / Sette A., Vitiello A., Reherman B., Fowler P., Nayersina R., Kast W.M., Melief C.J., Oseroff C., Yuan L., Ruppert J., Sidney J., Guercio M.F. del, Southwood S., Kubo R.T., Chesnut R.W., Grey H.M., Chisari F.V. // Journal of Immunology (Baltimore, Md.: 1950) – 1994. – Т. 153 – № 12 – С.5586–5592.

105. Karasev D.A. Application of molecular descriptors for recognition of phosphorylation sites in amino acid sequences / Karasev D.A., Savosina P.I., Sobolev B.N., Filimonov D.A., Lagunin A.A. // Biomeditsinskaya Khimiya – 2017. – Т. 63 – № 5 – С.423–427.

106. Filimonov D.A. Prediction of the Biological Activity Spectra of Organic Compounds Using the Pass Online Web Resource / Filimonov D.A., Lagunin A.A., Gloriozova T.A., Rudik A.V., Druzhilovskii D.S., Pogodin P.V., Poroikov V.V. // Chemistry of Heterocyclic Compounds – 2014. – Т. 50 – № 3 – С.444–457.

107. Landrum G. RDKit: Open-source cheminformatics. https://www.rdkit.org // – 2023.

108. Баскин И.И.Введение в хемоинформатику: моделирование «структура-свойство» / И. И. Баскин, Т. И. Маджидов, А. А. Варнек – Москва, Казань, Страсбург, 2020.

109. Hastie T.The Elements of Statistical Learning Data Mining, Inference, and Prediction / T. Hastie, R. Tibshirani, J. Friedman – Springer International Publishing, 2017. Вып. 2.

110. Youden W.J. Index for rating diagnostic tests / Youden W.J. // Cancer – 1950. – Т. 3 – № 1 – С.32–35.

111. Kuhn M. Building Predictive Models in *R* Using the **caret** Package / Kuhn M. // Journal of Statistical Software – 2008. – Т. 28 – № 5.

112. Pedregosa F. Scikit-learn: Machine Learning in Python / Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D. // MACHINE LEARNING IN PYTHON.

113. Chemoinformatics Approaches to Virtual Screening / / под ред. A. Varnek, A. Tropsha. – – The Royal Society of Chemistry, 2008.

114. Chang W. shiny: Web Application Framework for R // – 2023.

115. Specht G. Large database for the analysis and prediction of spliced and non-spliced peptide generation by proteasomes / Specht G., Roetschke H.P., Mansurkhodzhaev A., Henklein P., Textoris-Taube K., Urlaub H., Mishto M., Liepe J. // Scientific Data – 2020. – Т. 7 – № 1 – С.146.

116. Smirnov A.S. TCR‐Pred : A new web‐application for prediction of epitope and MHC specificity for CDR3 TCR sequences using molecular fragment descriptors / Smirnov A.S., Rudik A.V., Filimonov D.A., Lagunin A.A. // Immunology – 2023. – С.imm.13641.

# Приложение 1

Характеристика набор данных, используемых для предсказания сайтов расщепления белков протеасомой.

| **Набор данных** | **Радиус окна** | **Количество белков-источников** | **Количество пептидных фрагментов** | **Количество положительных примеров** | **Количество отрицательных примеров** |
| --- | --- | --- | --- | --- | --- |
| In vitro, 20S конститутивная протеасома | 2 | 88 | 2264 | 1384 | 1745 |
| 3 | 88 | 2264 | 1307 | 1646 |
| 4 | 88 | 2264 | 1228 | 1549 |
| In vitro, 20S, иммунопротеасома | 2 | 24 | 524 | 389 | 753 |
| 3 | 24 | 524 | 378 | 716 |
| 4 | 24 | 524 | 363 | 683 |
| In vivo | 2 | 56187 | 313811 | 204261 | 242055 |
| 3 | 56187 | 313811 | 271548 | 282171 |
| 4 | 56187 | 313811 | 273798 | 285122 |

# Приложение 2

Результаты внешней оценки качества модели прогноза связывания с HLA I с использованием оптимального уровня MNA дескрипторов методом пятикратной кросс-валидации. BA – balanced accuracy или сбалансированная точность. AP – average precision или средняя точность. Формулы расчета метрик указаны в тексте работы.

| **Аллель** | **Количество пептидов в общем датасете** | **AUC** | **Порог** | **Чувствительность** | **Специфичность** | **Точность** | **BA** | **AP** | **MCC** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| A\*01:03 | 7 | 0,99 | 0,37 | 1,00 | 0,98 | 0,98 | 0,99 | 0,00 | 0,02 |
| B\*51:08 | 624 | 0,98 | 0,00 | 0,97 | 0,96 | 0,96 | 0,97 | 0,07 | 0,15 |
| B\*39:06 | 1006 | 0,98 | 0,00 | 0,96 | 0,96 | 0,96 | 0,96 | 0,46 | 0,19 |
| A\*26:08 | 411 | 0,98 | 0,03 | 0,96 | 0,97 | 0,97 | 0,96 | 0,11 | 0,13 |
| B\*15:13 | 6 | 0,97 | 0,11 | 1,00 | 0,92 | 0,92 | 0,96 | 0,03 | 0,01 |
| B\*73:01 | 313 | 0,96 | 0,00 | 0,94 | 0,94 | 0,94 | 0,94 | 0,19 | 0,08 |
| B\*18:03 | 229 | 0,96 | 0,02 | 0,94 | 0,95 | 0,95 | 0,95 | 0,05 | 0,08 |
| B\*40:06 | 2494 | 0,96 | 0,02 | 0,94 | 0,95 | 0,95 | 0,94 | 0,19 | 0,24 |
| B\*41:01 | 1417 | 0,96 | 0,03 | 0,93 | 0,95 | 0,95 | 0,94 | 0,15 | 0,19 |
| B\*14:01 | 17 | 0,96 | 0,03 | 0,94 | 0,93 | 0,93 | 0,93 | 0,06 | 0,02 |
| B\*55:02 | 1575 | 0,96 | 0,00 | 0,94 | 0,94 | 0,94 | 0,94 | 0,11 | 0,18 |
| B\*83:01 | 40 | 0,96 | 0,01 | 0,95 | 0,91 | 0,91 | 0,93 | 0,01 | 0,02 |
| B\*57:02 | 80 | 0,96 | 0,02 | 0,95 | 0,93 | 0,93 | 0,94 | 0,01 | 0,04 |
| B\*50:01 | 1665 | 0,95 | 0,03 | 0,92 | 0,94 | 0,94 | 0,93 | 0,13 | 0,19 |
| B\*47:01 | 27 | 0,95 | 0,09 | 0,93 | 0,93 | 0,93 | 0,93 | 0,02 | 0,02 |
| A\*74:01 | 2697 | 0,95 | 0,04 | 0,91 | 0,95 | 0,95 | 0,93 | 0,20 | 0,24 |
| C\*14:03 | 2998 | 0,95 | 0,02 | 0,93 | 0,93 | 0,93 | 0,93 | 0,15 | 0,23 |
| A\*34:01 | 2386 | 0,95 | 0,02 | 0,92 | 0,94 | 0,94 | 0,93 | 0,15 | 0,21 |
| B\*38:02 | 3079 | 0,95 | 0,07 | 0,89 | 0,96 | 0,96 | 0,92 | 0,23 | 0,27 |
| B\*55:01 | 1744 | 0,95 | 0,01 | 0,93 | 0,93 | 0,93 | 0,93 | 0,11 | 0,17 |
| B\*27:02 | 3359 | 0,95 | 0,00 | 0,94 | 0,93 | 0,93 | 0,93 | 0,11 | 0,23 |
| B\*45:01 | 3781 | 0,94 | 0,06 | 0,89 | 0,95 | 0,95 | 0,92 | 0,36 | 0,27 |
| A\*66:01 | 2907 | 0,94 | 0,01 | 0,92 | 0,92 | 0,92 | 0,92 | 0,15 | 0,20 |
| B\*27:07 | 2127 | 0,94 | 0,00 | 0,94 | 0,93 | 0,93 | 0,93 | 0,06 | 0,19 |
| B\*27:08 | 2142 | 0,94 | 0,00 | 0,94 | 0,93 | 0,93 | 0,93 | 0,06 | 0,19 |
| B\*27:01 | 3889 | 0,94 | 0,00 | 0,93 | 0,93 | 0,93 | 0,93 | 0,13 | 0,25 |
| B\*40:01 | 14435 | 0,94 | 0,01 | 0,92 | 0,92 | 0,92 | 0,92 | 0,43 | 0,40 |
| B\*38:01 | 2815 | 0,94 | 0,06 | 0,89 | 0,94 | 0,94 | 0,91 | 0,24 | 0,22 |
| A\*11:02 | 2939 | 0,94 | 0,00 | 0,93 | 0,92 | 0,92 | 0,92 | 0,10 | 0,20 |
| B\*35:03 | 3052 | 0,94 | 0,00 | 0,92 | 0,91 | 0,91 | 0,92 | 0,12 | 0,20 |
| B\*56:01 | 2281 | 0,94 | 0,03 | 0,90 | 0,93 | 0,93 | 0,91 | 0,14 | 0,19 |
| A\*25:01 | 2760 | 0,94 | 0,05 | 0,89 | 0,93 | 0,93 | 0,91 | 0,26 | 0,21 |
| A\*24:07 | 1465 | 0,94 | 0,01 | 0,92 | 0,92 | 0,92 | 0,92 | 0,05 | 0,14 |
| B\*44:03 | 7421 | 0,94 | 0,00 | 0,92 | 0,92 | 0,92 | 0,92 | 0,23 | 0,31 |
| A\*33:03 | 5566 | 0,94 | 0,04 | 0,90 | 0,93 | 0,93 | 0,91 | 0,24 | 0,28 |
| B\*44:28 | 119 | 0,94 | 0,01 | 0,92 | 0,91 | 0,91 | 0,91 | 0,01 | 0,04 |
| B\*27:06 | 1280 | 0,94 | 0,00 | 0,92 | 0,92 | 0,92 | 0,92 | 0,04 | 0,13 |
| C\*08:02 | 5586 | 0,93 | 0,03 | 0,90 | 0,91 | 0,91 | 0,91 | 0,23 | 0,26 |
| B\*35:07 | 1953 | 0,93 | 0,05 | 0,89 | 0,93 | 0,93 | 0,91 | 0,08 | 0,17 |
| B\*27:04 | 1348 | 0,93 | 0,00 | 0,92 | 0,92 | 0,92 | 0,92 | 0,04 | 0,14 |
| B\*07:04 | 1494 | 0,93 | 0,00 | 0,91 | 0,91 | 0,91 | 0,91 | 0,06 | 0,14 |
| B\*49:01 | 5872 | 0,93 | 0,07 | 0,87 | 0,93 | 0,93 | 0,90 | 0,33 | 0,28 |
| A\*33:01 | 3848 | 0,93 | 0,00 | 0,91 | 0,91 | 0,91 | 0,91 | 0,17 | 0,21 |
| B\*44:02 | 11785 | 0,93 | 0,01 | 0,90 | 0,91 | 0,91 | 0,91 | 0,39 | 0,35 |
| A\*02:17 | 521 | 0,93 | 0,03 | 0,90 | 0,90 | 0,90 | 0,90 | 0,04 | 0,08 |
| B\*08:02 | 56 | 0,93 | 0,06 | 0,88 | 0,94 | 0,94 | 0,91 | 0,01 | 0,03 |
| A\*03:01 | 15427 | 0,93 | 0,01 | 0,91 | 0,91 | 0,91 | 0,91 | 0,34 | 0,39 |
| B\*53:01 | 2990 | 0,93 | 0,00 | 0,91 | 0,90 | 0,90 | 0,91 | 0,13 | 0,18 |
| B\*27:03 | 1040 | 0,93 | 0,00 | 0,91 | 0,91 | 0,91 | 0,91 | 0,03 | 0,11 |
| B\*27:10 | 12 | 0,93 | 0,15 | 0,92 | 0,91 | 0,91 | 0,91 | 0,01 | 0,01 |
| A\*68:01 | 9452 | 0,93 | 0,00 | 0,91 | 0,9 | 0,90 | 0,90 | 0,26 | 0,30 |
| B\*35:08 | 1214 | 0,93 | 0,05 | 0,87 | 0,91 | 0,91 | 0,89 | 0,15 | 0,12 |
| B\*39:01 | 2228 | 0,93 | 0,02 | 0,88 | 0,90 | 0,90 | 0,89 | 0,13 | 0,15 |
| B\*52:01 | 1904 | 0,93 | 0,09 | 0,85 | 0,94 | 0,94 | 0,90 | 0,21 | 0,18 |
| B\*44:27 | 208 | 0,93 | 0,08 | 0,87 | 0,93 | 0,93 | 0,90 | 0,01 | 0,05 |
| B\*18:01 | 5568 | 0,93 | 0,05 | 0,87 | 0,92 | 0,92 | 0,89 | 0,38 | 0,25 |
| B\*40:02 | 17116 | 0,93 | 0,00 | 0,91 | 0,90 | 0,90 | 0,91 | 0,32 | 0,40 |
| B\*15:03 | 4000 | 0,92 | 0,06 | 0,86 | 0,92 | 0,92 | 0,89 | 0,42 | 0,22 |
| B\*42:01 | 3539 | 0,92 | 0,00 | 0,91 | 0,90 | 0,90 | 0,91 | 0,09 | 0,20 |
| B\*14:02 | 2288 | 0,92 | 0,07 | 0,85 | 0,92 | 0,92 | 0,89 | 0,27 | 0,17 |
| B\*39:24 | 396 | 0,92 | 0,01 | 0,89 | 0,90 | 0,90 | 0,89 | 0,03 | 0,06 |
| B\*13:02 | 4371 | 0,92 | 0,00 | 0,89 | 0,89 | 0,89 | 0,89 | 0,23 | 0,20 |
| A\*34:02 | 3806 | 0,92 | 0,00 | 0,89 | 0,89 | 0,89 | 0,89 | 0,15 | 0,19 |
| B\*44:05 | 16 | 0,92 | 0,18 | 0,81 | 0,96 | 0,96 | 0,89 | 0,17 | 0,02 |
| A\*11:01 | 14082 | 0,92 | 0,01 | 0,89 | 0,89 | 0,89 | 0,89 | 0,30 | 0,34 |
| C\*15:02 | 4474 | 0,92 | 0,04 | 0,87 | 0,90 | 0,90 | 0,88 | 0,19 | 0,20 |
| C\*04:03 | 1142 | 0,92 | 0,06 | 0,86 | 0,91 | 0,91 | 0,88 | 0,07 | 0,11 |
| B\*58:02 | 1105 | 0,92 | 0,07 | 0,85 | 0,91 | 0,91 | 0,88 | 0,07 | 0,11 |
| B\*07:01 | 60 | 0,91 | 0,08 | 0,9 | 0,91 | 0,91 | 0,90 | 0,02 | 0,03 |
| B\*54:01 | 1882 | 0,91 | 0,05 | 0,86 | 0,90 | 0,90 | 0,88 | 0,11 | 0,14 |
| C\*01:02 | 2703 | 0,91 | 0,07 | 0,84 | 0,90 | 0,90 | 0,87 | 0,17 | 0,16 |
| A\*26:01 | 2832 | 0,91 | 0,04 | 0,86 | 0,89 | 0,89 | 0,87 | 0,14 | 0,16 |
| B\*13:01 | 4147 | 0,91 | 0,00 | 0,88 | 0,87 | 0,87 | 0,88 | 0,16 | 0,18 |
| A\*29:02 | 8905 | 0,91 | 0,00 | 0,88 | 0,87 | 0,87 | 0,88 | 0,24 | 0,26 |
| A\*31:01 | 5768 | 0,91 | 0,00 | 0,89 | 0,88 | 0,88 | 0,88 | 0,13 | 0,22 |
| A\*02:07 | 4783 | 0,91 | 0,01 | 0,88 | 0,87 | 0,87 | 0,88 | 0,12 | 0,19 |
| B\*07:02 | 23920 | 0,91 | 0,05 | 0,85 | 0,89 | 0,89 | 0,87 | 0,41 | 0,41 |
| C\*03:02 | 1277 | 0,91 | 0,03 | 0,87 | 0,87 | 0,87 | 0,87 | 0,03 | 0,10 |
| A\*01:01 | 14460 | 0,90 | 0,03 | 0,86 | 0,88 | 0,88 | 0,87 | 0,37 | 0,31 |
| A\*24:02 | 15190 | 0,90 | 0,09 | 0,84 | 0,90 | 0,90 | 0,87 | 0,28 | 0,35 |
| C\*08:01 | 1954 | 0,90 | 0,02 | 0,87 | 0,87 | 0,87 | 0,87 | 0,04 | 0,12 |
| B\*58:01 | 6526 | 0,90 | 0,05 | 0,85 | 0,89 | 0,89 | 0,87 | 0,17 | 0,22 |
| C\*17:01 | 1505 | 0,90 | 0,03 | 0,86 | 0,86 | 0,86 | 0,86 | 0,05 | 0,10 |
| B\*35:02 | 38 | 0,90 | 0,16 | 0,82 | 0,96 | 0,96 | 0,89 | 0,00 | 0,03 |
| A\*23:01 | 4544 | 0,90 | 0,11 | 0,82 | 0,92 | 0,92 | 0,87 | 0,12 | 0,22 |
| A\*24:13 | 250 | 0,90 | 0,10 | 0,84 | 0,90 | 0,90 | 0,87 | 0,01 | 0,05 |
| C\*04:01 | 3791 | 0,90 | 0,03 | 0,86 | 0,86 | 0,86 | 0,86 | 0,10 | 0,15 |
| B\*51:01 | 8854 | 0,90 | 0,10 | 0,81 | 0,90 | 0,90 | 0,86 | 0,31 | 0,27 |
| B\*27:20 | 89 | 0,90 | 0,09 | 0,83 | 0,92 | 0,92 | 0,88 | 0,01 | 0,03 |
| B\*15:10 | 3679 | 0,90 | 0,01 | 0,86 | 0,85 | 0,85 | 0,85 | 0,15 | 0,15 |
| A\*29:01 | 144 | 0,90 | 0,09 | 0,83 | 0,90 | 0,90 | 0,87 | 0,00 | 0,04 |
| B\*35:01 | 6602 | 0,89 | 0,03 | 0,85 | 0,86 | 0,86 | 0,85 | 0,15 | 0,20 |
| B\*15:01 | 15163 | 0,89 | 0,01 | 0,85 | 0,86 | 0,86 | 0,86 | 0,28 | 0,29 |
| B\*15:17 | 2276 | 0,89 | 0,06 | 0,83 | 0,88 | 0,88 | 0,86 | 0,06 | 0,13 |
| B\*37:01 | 2034 | 0,89 | 0,16 | 0,77 | 0,93 | 0,93 | 0,85 | 0,28 | 0,15 |
| A\*02:04 | 3527 | 0,89 | 0,00 | 0,87 | 0,86 | 0,86 | 0,86 | 0,07 | 0,15 |
| B\*57:01 | 19469 | 0,89 | 0,05 | 0,83 | 0,87 | 0,87 | 0,85 | 0,36 | 0,33 |
| B\*46:01 | 2910 | 0,89 | 0,01 | 0,86 | 0,85 | 0,85 | 0,86 | 0,06 | 0,13 |
| B\*57:03 | 6480 | 0,89 | 0,00 | 0,86 | 0,85 | 0,85 | 0,86 | 0,11 | 0,20 |
| A\*02:03 | 5678 | 0,89 | 0,00 | 0,87 | 0,86 | 0,86 | 0,86 | 0,08 | 0,19 |
| A\*24:03 | 336 | 0,89 | 0,01 | 0,85 | 0,85 | 0,85 | 0,85 | 0,01 | 0,05 |
| A\*30:04 | 100 | 0,88 | 0,15 | 0,80 | 0,93 | 0,93 | 0,86 | 0,02 | 0,03 |
| B\*40:13 | 52 | 0,88 | 0,14 | 0,81 | 0,90 | 0,90 | 0,85 | 0,00 | 0,02 |
| A\*24:06 | 407 | 0,88 | 0,11 | 0,81 | 0,89 | 0,89 | 0,85 | 0,02 | 0,06 |
| B\*15:08 | 29 | 0,88 | 0,15 | 0,79 | 0,93 | 0,93 | 0,86 | 0,00 | 0,02 |
| A\*36:01 | 2713 | 0,88 | 0,01 | 0,84 | 0,84 | 0,84 | 0,84 | 0,11 | 0,12 |
| A\*26:03 | 66 | 0,88 | 0,07 | 0,85 | 0,86 | 0,86 | 0,85 | 0,00 | 0,02 |
| A\*30:01 | 3907 | 0,88 | 0,03 | 0,84 | 0,85 | 0,85 | 0,84 | 0,10 | 0,15 |
| B\*44:08 | 49 | 0,88 | 0,02 | 0,86 | 0,83 | 0,83 | 0,84 | 0,00 | 0,02 |
| B\*08:01 | 8947 | 0,88 | 0,06 | 0,81 | 0,87 | 0,86 | 0,84 | 0,32 | 0,22 |
| C\*05:01 | 11120 | 0,88 | 0,05 | 0,82 | 0,85 | 0,85 | 0,84 | 0,21 | 0,24 |
| A\*80:01 | 159 | 0,88 | 0,03 | 0,84 | 0,87 | 0,87 | 0,85 | 0,00 | 0,03 |
| A\*02:05 | 4083 | 0,88 | 0,02 | 0,84 | 0,84 | 0,84 | 0,84 | 0,08 | 0,15 |
| C\*03:04 | 7559 | 0,88 | 0,01 | 0,85 | 0,83 | 0,83 | 0,84 | 0,11 | 0,19 |
| A\*68:23 | 76 | 0,87 | 0,01 | 0,86 | 0,82 | 0,82 | 0,84 | 0,00 | 0,02 |
| A\*02:02 | 5994 | 0,87 | 0,00 | 0,85 | 0,84 | 0,84 | 0,85 | 0,08 | 0,18 |
| B\*39:05 | 8 | 0,87 | 0,21 | 0,75 | 1,00 | 1,00 | 0,87 | 0,06 | 0,29 |
| A\*69:01 | 1044 | 0,87 | 0,13 | 0,77 | 0,88 | 0,88 | 0,83 | 0,04 | 0,08 |
| A\*30:02 | 3871 | 0,87 | 0,10 | 0,78 | 0,88 | 0,88 | 0,83 | 0,10 | 0,16 |
| B\*27:09 | 7681 | 0,87 | 0,07 | 0,82 | 0,86 | 0,86 | 0,84 | 0,10 | 0,21 |
| C\*07:02 | 2886 | 0,87 | 0,08 | 0,80 | 0,86 | 0,86 | 0,83 | 0,09 | 0,13 |
| A\*32:07 | 78 | 0,87 | 0,07 | 0,82 | 0,84 | 0,84 | 0,83 | 0,00 | 0,02 |
| A\*02:11 | 2633 | 0,87 | 0,00 | 0,85 | 0,83 | 0,83 | 0,84 | 0,04 | 0,11 |
| C\*07:06 | 68 | 0,87 | 0,03 | 0,85 | 0,81 | 0,81 | 0,83 | 0,00 | 0,02 |
| A\*32:15 | 59 | 0,87 | 0,23 | 0,76 | 0,93 | 0,93 | 0,85 | 0,00 | 0,03 |
| C\*14:02 | 5418 | 0,87 | 0,06 | 0,81 | 0,83 | 0,83 | 0,82 | 0,11 | 0,15 |
| A\*02:06 | 4926 | 0,86 | 0,01 | 0,85 | 0,83 | 0,83 | 0,84 | 0,06 | 0,16 |
| C\*16:01 | 5579 | 0,86 | 0,07 | 0,80 | 0,85 | 0,85 | 0,82 | 0,11 | 0,16 |
| A\*26:02 | 174 | 0,86 | 0,00 | 0,82 | 0,83 | 0,83 | 0,82 | 0,01 | 0,03 |
| A\*02:19 | 240 | 0,86 | 0,02 | 0,82 | 0,83 | 0,83 | 0,83 | 0,01 | 0,03 |
| A\*02:16 | 182 | 0,86 | 0,02 | 0,86 | 0,80 | 0,80 | 0,83 | 0,00 | 0,03 |
| A\*02:50 | 88 | 0,86 | 0,02 | 0,84 | 0,82 | 0,82 | 0,83 | 0,00 | 0,02 |
| B\*81:01 | 22 | 0,85 | 0,02 | 0,82 | 0,82 | 0,82 | 0,82 | 0,05 | 0,01 |
| A\*68:02 | 6280 | 0,85 | 0,12 | 0,76 | 0,86 | 0,86 | 0,81 | 0,14 | 0,17 |
| B\*41:06 | 22 | 0,85 | 0,12 | 0,82 | 0,85 | 0,85 | 0,83 | 0,00 | 0,01 |
| B\*15:09 | 75 | 0,85 | 0,00 | 0,85 | 0,77 | 0,77 | 0,81 | 0,00 | 0,02 |
| C\*02:02 | 3104 | 0,85 | 0,08 | 0,78 | 0,84 | 0,84 | 0,81 | 0,05 | 0,12 |
| B\*15:11 | 181 | 0,85 | 0,03 | 0,81 | 0,81 | 0,81 | 0,81 | 0,00 | 0,03 |
| C\*15:05 | 110 | 0,85 | 0,06 | 0,78 | 0,85 | 0,85 | 0,82 | 0,00 | 0,02 |
| A\*32:01 | 5689 | 0,85 | 0,05 | 0,79 | 0,82 | 0,82 | 0,80 | 0,11 | 0,15 |
| A\*66:02 | 31 | 0,85 | 0,22 | 0,71 | 0,92 | 0,92 | 0,82 | 0,01 | 0,02 |
| B\*15:02 | 20708 | 0,85 | 0,03 | 0,80 | 0,81 | 0,81 | 0,80 | 0,26 | 0,26 |
| C\*12:03 | 5097 | 0,84 | 0,01 | 0,82 | 0,79 | 0,79 | 0,81 | 0,06 | 0,13 |
| C\*03:03 | 4602 | 0,84 | 0,03 | 0,80 | 0,81 | 0,81 | 0,81 | 0,06 | 0,13 |
| B\*48:01 | 116 | 0,84 | 0,03 | 0,80 | 0,81 | 0,81 | 0,81 | 0,00 | 0,02 |
| B\*27:05 | 55191 | 0,84 | 0,14 | 0,74 | 0,86 | 0,85 | 0,80 | 0,44 | 0,43 |
| A\*02:01 | 52723 | 0,84 | 0,11 | 0,77 | 0,84 | 0,84 | 0,80 | 0,38 | 0,41 |
| A\*02:12 | 327 | 0,84 | 0,01 | 0,80 | 0,80 | 0,80 | 0,80 | 0,00 | 0,03 |
| C\*07:01 | 4807 | 0,84 | 0,15 | 0,73 | 0,86 | 0,86 | 0,79 | 0,06 | 0,14 |
| C\*07:04 | 1284 | 0,84 | 0,07 | 0,77 | 0,82 | 0,82 | 0,80 | 0,02 | 0,07 |
| C\*12:02 | 11189 | 0,83 | 0,06 | 0,77 | 0,81 | 0,81 | 0,79 | 0,14 | 0,19 |
| B\*40:32 | 7 | 0,83 | 0,42 | 0,71 | 0,97 | 0,97 | 0,84 | 0,01 | 0,01 |
| C\*12:04 | 72 | 0,83 | 0,00 | 0,78 | 0,80 | 0,80 | 0,79 | 0,00 | 0,02 |
| B\*44:09 | 212 | 0,82 | 0,07 | 0,76 | 0,81 | 0,81 | 0,79 | 0,00 | 0,03 |
| B\*08:03 | 23 | 0,81 | 0,07 | 0,74 | 0,83 | 0,83 | 0,79 | 0,00 | 0,01 |
| A\*03:02 | 30 | 0,81 | 0,04 | 0,77 | 0,80 | 0,80 | 0,78 | 0,00 | 0,01 |
| C\*16:02 | 10 | 0,81 | 0,47 | 0,70 | 0,97 | 0,97 | 0,83 | 0,00 | 0,01 |
| B\*41:04 | 52 | 0,81 | 0,00 | 0,79 | 0,77 | 0,77 | 0,78 | 0,00 | 0,01 |
| C\*06:02 | 4481 | 0,81 | 0,15 | 0,70 | 0,83 | 0,83 | 0,77 | 0,08 | 0,12 |
| B\*07:06 | 15 | 0,79 | 0,07 | 0,73 | 0,82 | 0,82 | 0,78 | 0,01 | 0,01 |
| A\*03:19 | 14 | 0,79 | 0,16 | 0,71 | 0,83 | 0,83 | 0,77 | 0,00 | 0,01 |
| A\*30:03 | 20 | 0,78 | 0,77 | 0,50 | 1,00 | 1,00 | 0,75 | 0,14 | 0,38 |
| B\*51:02 | 7 | 0,78 | 0,03 | 0,71 | 0,82 | 0,82 | 0,77 | 0,01 | 0,00 |
| B\*42:02 | 15 | 0,78 | 0,30 | 0,60 | 0,93 | 0,93 | 0,77 | 0,01 | 0,01 |
| B\*14:03 | 21 | 0,78 | 0,41 | 0,62 | 0,91 | 0,91 | 0,76 | 0,00 | 0,01 |
| B\*41:03 | 85 | 0,76 | 0,06 | 0,69 | 0,77 | 0,77 | 0,73 | 0,00 | 0,01 |
| A\*11:10 | 6 | 0,75 | 0,07 | 0,75 | 0,82 | 0,82 | 0,79 | 0,00 | 0,00 |
| C\*03:01 | 88 | 0,75 | 0,04 | 0,69 | 0,74 | 0,74 | 0,72 | 0,00 | 0,01 |
| B\*15:16 | 19 | 0,75 | 0,43 | 0,53 | 0,92 | 0,92 | 0,72 | 0,00 | 0,01 |
| B\*41:02 | 21 | 0,73 | 0,47 | 0,52 | 0,89 | 0,89 | 0,71 | 0,00 | 0,01 |
| B\*39:09 | 21 | 0,71 | 0,46 | 0,48 | 0,94 | 0,94 | 0,71 | 0,00 | 0,01 |