НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2023 Том 23 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2023 Vol. 23 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

doi: 10.17586/2226-1494-2023-23-5-989-1000
УДК 004.94

# Метод построения интерпретируемых скрытых марковских моделей для задачи поиска связываемых участков пептидов в последовательностях белков

**Денис Анатольевич Клеверов[1]✉, Анатолий Абрамович Шалыто[2], Максим Артемов[3]**

[1,2,3] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
[3] Университет Вашингтона в Сент-Луисе. Медицинская Школа. Отдел патологии и иммунологии, Сент-Луис, 63110, США

[1] denklewer@gmail.com✉, https://orcid.org/0009-0002-1362-486X
[2] shalyto@mail.ifmo.ru, https://orcid.org/0000-0002-2723-2077
[3] martyomov@pathology.wustl.edu, https://orcid.org/0000-0002-1133-4212

**Аннотация**

**Введение.** Решение задачи предсказания иммунного ответа организма на чужеродные фрагменты белковых последовательностей, обработанные клеткой, является ключевым этапом разработки персонализированных вакцин от рака. Отбор пептидов, участвующих в иммунном ответе, представляет собой сложный многоступенчатый процесс фильтрации исходных последовательностей для презентации их фрагментов на поверхности клетки. Наиболее изученной является задача предсказания одного из этапов такой фильтрации — вероятности связывания пептидов с молекулами главного комплекса гистосовместимости. Современные методы предсказания данного этапа обычно основаны на алгоритмах, использующих искусственные нейронные сети, что не позволяет в должной мере интерпретировать результаты работы моделей. Одним из методов решения проблемы является использование интерпретируемых скрытых марковских моделей. В работе выполнен анализ задачи предсказания связывающей способности и предложен метод построения интерпретируемых моделей, учитывающих ограничения и требования предметной области. **Метод.** Разработан метод построения, обучения и интерпретации скрытых марковских моделей для каждого класса молекул. Построение и обучение моделей основано на поддержании архитектуры модели, способной извлекать и визуализировать связываемый участок пептида. Интерпретация возможна благодаря анализу графа модели. **Основные результаты.** Предложенный метод протестирован в задаче обучения модели, позволяющей помимо предсказания получать позицию связываемого участка пептида и распределение аминокислот в нем. Обучены модели предсказания для двух разновидностей молекул с использованием данных связывания. Распределения аминокислот связываемого участка совпадают с распределениями состояний модели. Паттерны последовательностей участков, извлеченные с помощью обученных моделей для двух наборов пептидных данных, соответствуют паттернам из открытых источников, что подтверждает успешную апробацию метода. **Обсуждение.** Интерпретируемые модели лучше описывают предметную область задачи и помогают сделать выводы о характеристиках пептидов, основываясь на информации, извлеченной из модели. Эта информация позволит исследователям лучше понять остальные шаги процессинга пептидов при иммунном ответе: изучить взаимосвязи между ними и произвести перенос знаний из моделей, обученных для одного этапа, на другие. Таким образом, предлагаемый метод построения позволит обучать модели в условиях недостатка обучающих данных.

**Ключевые слова**

предсказание связывающей способности, скрытые марковские модели, алгоритм Витерби, анализ данных, поиск мотива, выравнивание последовательностей

**Ссылка для цитирования:** Клеверов Д.А., Шалыто А.А., Артемов М. Метод построения интерпретируемых скрытых марковских моделей для задачи поиска связываемых участков пептидов в последовательностях белков // Научно-технический вестник информационных технологий, механики и оптики. 2023. Т. 23, № 5. С. 989–1000. doi: 10.17586/2226-1494-2023-23-5-989-1000

# A method for constructing interpretable hidden Markov models for the task of identifying binding cores in sequences

**Denis A. Kleverov[1]✉, Anatoly A. Shalyto[2], Maxim N. Artyomov[3]**

[1,2,3] ITMO University, Saint Petersburg, 197101, Russian Federation

[3] Washington University in St. Louis. School of Medicine. Department of Pathology and Immunology, Saint Louis, 63110, USA

[1] denklewer@gmail.com✉, https://orcid.org/0009-0002-1362-486X

[2] shalyto@mail.ifmo.ru, https://orcid.org/0000-0002-2723-2077

[3] martyomov@pathology.wustl.edu, https://orcid.org/0000-0002-1133-4212

**Abstract**

Solving the problem of predicting the immune response against foreign protein sequence fragments processed by cells is one of the major milestones on the road to the personalized cancer vaccine development. The selection of peptides participating in the immune response is a complex multi-stage process of filtering initial sequences to present their fragments on the cell surface. The most studied task regarding this filtering nowadays is the prediction of the binding probability of peptides to major histocompatibility complex molecules. Modern methods for predicting this stage are usually based on algorithms using artificial neural networks, which make it impossible to interpret the result predictions of such models. One of the methods to overcome this limitation is the use of interpretable hidden Markov models. In this work, an analysis of the binding prediction task is performed. As a result, a method for constructing interpretable models that consider domain-specific constraints and requirements is proposed. A method for the constriction, training and interpretation of hidden Markov models was proposed for each class of molecules. The construction and training are based on maintaining the model architecture capable of extracting and visualizing the binding core of the peptide. Interpretation is possible through the analysis of the model graph. The proposed method is tested in the task of training a model that not only enables prediction but also facilitates determining the position of the peptide binding core and the distribution of amino acids within the core. Prediction models were trained for two types of molecules using binding data. The distributions of amino acids in the binding core match the state distributions of the model. Sequence patterns of such regions extracted using the trained models for two sets of peptide data correspond to patterns from public databases, confirming the successful validation of the method. Interpretable models provide a better description of the problem domain and help to draw a conclusion about peptide characteristics based on information extracted from the model. This information will allow researchers to better understand other steps of peptide processing involved in the immune response. For example, one can study relationships between these steps or perform a transfer of knowledge from models trained for one step to others. Using this knowledge will allow the training of the models under conditions of limited training data.

**Keywords**

binding prediction, hidden Markov models, Viterbi algorithm, data analysis, motif identification, sequences alignment, interpretable models

**For citation:** Kleverov D.A., Shalyto A.A., Artyomov M.N. A method for constructing interpretable hidden Markov models for the task of identifying binding cores in sequences. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2023, vol. 23, no. 5, pp. 989–1000 (in Russian). doi: 10.17586/2226-1494-2023-23-5-989-1000

## Введение

Формирование иммунного ответа к патогенам в организме представляет собой сложный процесс, состоящий из множества связанных между собой факторов [1]. При этом наиболее значимой является процедура процессинга белковых последовательностей (пептидов). Процедура процессинга — основной механизм запуска иммунных реакций, содержащий процесс поэтапного отбора клеткой фрагментов внутренних и внешних антигенов для их презентации на поверхности клеточной мембраны посредством связывания с молекулами главного комплекса гистосовместимости (ГКГС). Этот комплекс представляет собой белки, оборудованные специальным рецептором для размещения пептидов и располагающиеся на поверхности клеток. В последствии связанные фрагменты (эпитопы) «инспектируются» иммунными клетками, способными отличать «свои» последовательности от «чужих». Именно результат этого распознавания запускает процесс иммунной реакции, что приводит к формированию адаптивного иммунного ответа на враждебный патоген [2].

Таким образом, генерация иммунного ответа — результат многоступенчатого процесса [3], состоящего из поглощения антигена клеткой организма, получения пептида посредством обработки протеасомой, которая разделяет белки на короткие фрагменты, транспортировки пептида к молекуле ГКГС, связывания пептида с молекулой (формирование комплекса), доставки результирующего комплекса на поверхность клетки и его распознавания Т-клеточным рецептором.

Нахождение белковых последовательностей, участвующих в описанной генерации иммунного ответа (неоэпитопов), является одним из важнейших шагов в процессе разработки вакцины от рака [4]. Пептиды, имеющие высокий шанс быть презентованными и распознанными иммунными клетками, могут быть использованы в качестве компонентов персонализированных вакцинаций [5] для стимуляции иммунного ответа даже в тех случаях, когда иммунная реакция не запускается автоматически [6].

С практической точки зрения отбор пептидов-кандидатов для вакцинации может быть сведен к задаче построения предсказательной модели. Такая модель по входной белковой последовательности выдает предска-

зание, представляющее собой численную характеристику, отражающую успешность прохождения того или иного этапа процессинга данным фрагментом. Белковая последовательность является последовательностью символов неизвестной длины, где каждый символ соответствует общепринятому сокращению названия одной из 20 аминокислот (ACDEFGHIKLMNPQRSTVWY). Например, A — аланин, а Y — тирозин.

Современные методы отбора пептидов [7, 8] основаны на применении нейронных сетей к задаче предсказания связывающей способности пептида по отношению к ГКГС и презентации результирующего комплекса на поверхности клетки. Однако эти методы способны обрабатывать только пептиды фиксированной длины. Решением этой проблемы стало использование рекуррентных нейронных сетей [9, 10]. Однако их основным недостатком является сложность интерпретации результатов и переноса знаний на другие модели.

Решением для обработки последовательностей различных длин также является использование скрытых марковских моделей [11, 12]. Эти модели уже показали свою эффективность при решении задачи предсказания связывающей способности [6]. При этом важным достоинством таких моделей является возможность интерпретации и визуализации их графов [13]. Однако в настоящее время отсутствуют методы построения и анализа подобных моделей, которые учитывали бы все особенности поставленной задачи.

В данной работе предложен метод построения, обучения и анализа интерпретируемых марковских моделей для предсказания связывающей способности пептидов с учетом особенностей этой задачи. Такие модели, кроме предсказания факта связывания, способны находить связываемый участок внутри анализируемого пептида, что помогает добиться интерпретируемости модели.

## Особенности задачи предсказания и их влияние на данные

Благодаря своей биологической природе, рассматриваемая задача имеет ряд особенностей, которые исследователю необходимо учитывать в процессе построения модели.

**Объем доступных данных для разных этапов.** В зависимости от этапа процессинга объем данных, доступных для обучения модели, отличается. С развитием масс-спектрометрии объем доступных данных связывания и презентации существенно увеличился [14, 15]. Однако известно лишь несколько сотен пептидов для этапов расщепления протеина на фрагменты [16] и распознавания Т-клетками, что не позволяет построить достаточно адекватную модель для этих этапов, используя только указанные данные [17]. Однако тот факт, что в ходе всей процедуры процессинга происходит поэтапное отсеивание последовательностей пептидов, позволяет предположить, что модель, построенная для одного этапа, может быть использована и для лучшего понимания других этапов [18]. Переиспользование готовых моделей дало бы возможность существенно сократить область поиска модели за счет ограничения пространства всех пептидов теми из них, что прошли первые этапы отбора. Это может позволить построить статистически значимую модель.

**Структура молекул ГКГС.** Другой особенностью задачи предсказания являются различия в физической структуре связывающих молекул ГКГС. Эти различия, в свою очередь, влекут изменения в структуре последовательности пептида. Различные позиции внутри пептида по-разному влияют на каждый этап процессинга [18]. Например, в задаче предсказания связывающей способности фрагмент, играющий непосредственную роль в связывании, является наиболее важным [19]. Он называется связываемым участком.

Молекулы ГКГС делятся на два класса: первый (I) и второй (II).

Связывающий участок молекул класса I состоит из двух спиралей, формирующих закрытый связывающий «карман» (рис. 1). Это позволяет молекуле связывать пептиды небольших длин (обычно от 9 до 12) таким образом, что важными позициями в этом взаимодействии являются аминокислоты на концах пептида. Отметим, что центральная часть пептида может не участвовать во взаимодействии или участвовать частично [20].

Связывающий «карман» для класса II молекул открыт, что позволяет концам пептида «свисать» (рис. 1) за его пределы. Молекула в этом случае способна связывать пептиды более широкого спектра длин, чем молекула класса I (обычно от 12 до 25). Важным в этом взаимодействии по-прежнему является более короткий связываемый участок, располагающийся с некоторым сдвигом для каждой последовательности.

При этом внутри связывающего участка также выделяют аминокислоты, непосредственно участвующие в связывании, которые называются якорями.

Таким образом, особенности структур молекул и их взаимодействие с пептидами приводит к постановке задачи поиска участка пептида, участвующего во взаимодействии. В ходе решения этой задачи последовательности различной длины должны быть выравнены по отношению друг к другу. Под выравниванием понимается выравнивание последовательностей в биологическом смысле, где схожие элементы последовательностей совмещаются друг с другом. Результирующий паттерн последовательностей, полученный таким образом, используется при дальнейшем анализе. Процесс выравнивания состоит в совмещении характерных аминокислот связываемого участка для различных пептидов одной молекулы. При этом пептиды для молекул класса I выравниваются совмещением концов последовательностей пептидов и пропуском аминокислот для более длинных пептидов в середине. Для молекул класса II совмещение фрагментов связываемых участков пептидов происходит с той особенностью, что возможны различные сдвиги целого участка внутри пептида. Без процедуры выравнивания основной мотив (паттерн последовательностей) связываемых фрагментов найти невозможно.

Например, на рис. 1 приведены схематичные изображения молекул и логотипы главных паттернов для классов I (рис. 1, *a*, *b*) и II (рис. 1, *c*, *d*) для длин пептидов 9 и 15 соответственно. Логотип — графическое

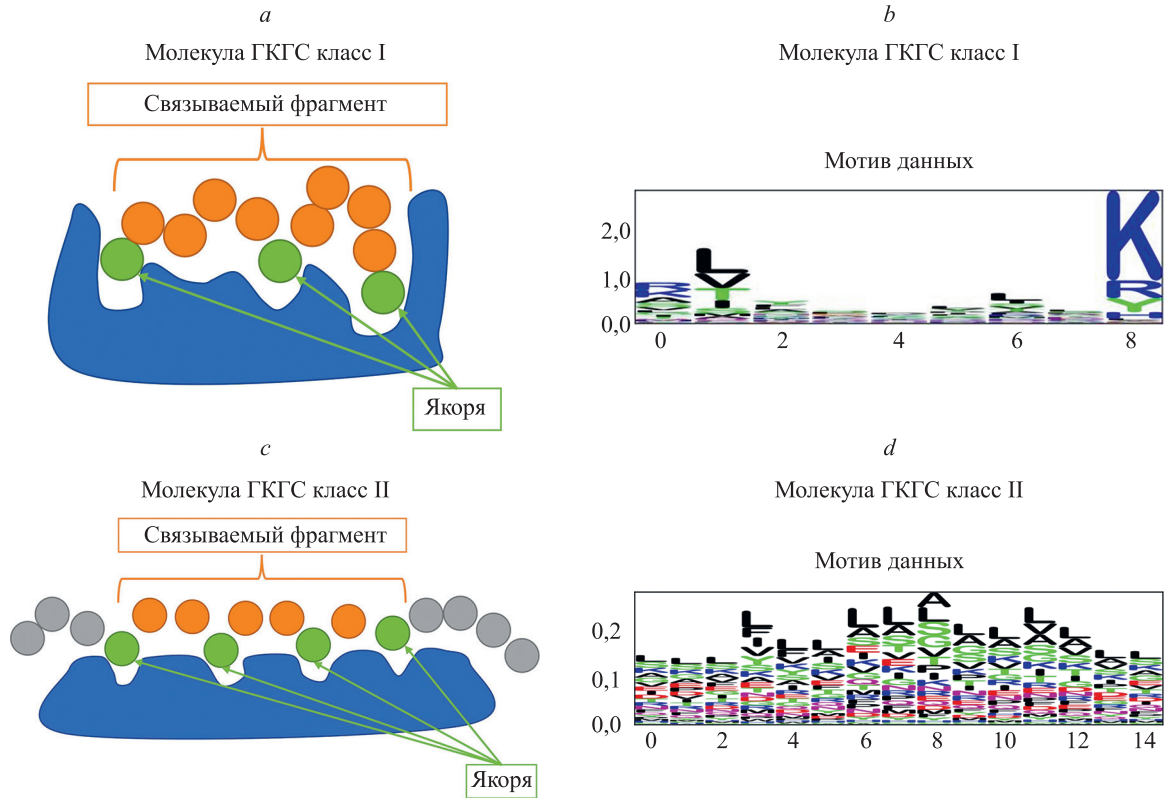



*Рис. 1*. Схемы связывания для молекул классов I (*a*) и II (*c*) и соответствующие схемам паттерны последовательностей классов I (*b*) и II (*d*)

*Fig. 1*. Binding schematics for class I (*a*) and class II (*c*) molecules, alongside motif patterns for the respective schematics for the class I (*b*) and class II (*d*)

представление паттерна набора последовательностей, в котором по оси $X$ отложены позиции аминокислот пептида, а по оси $Y$ — информационное содержание аминокислот в битах. Размер символов отображает частотность аминокислоты на данной позиции последовательности. На полученном графике видно, что для изначальных последовательностей молекулы класса II какой-либо закономерности в данных не наблюдается.

**Разнообразие аллелей молекул ГКГС.** Одним из важнейших свойств молекул ГКГС является их разнообразие. Внутри одного организма может быть до шести видов (аллелей) молекул каждого класса, однако внутри популяции эти молекулы чрезвычайно разнообразны [21]. В настоящее время известно более 20 000 вариаций молекул класса I и более 8000 вариаций для класса II [22].

При этом каждая разновидность (аллель) такой молекулы уникальна по своим свойствам и способна связываться только с пептидами определенного вида, что порождает уникальный мотив связывания для каждой аллели. На рис. 2 приведены такие мотивы для четырех молекул класса I, по которым можно судить о различиях предпочитаемых этими молекулами пептидов. Например, молекула HLA-A*03:01 предпочитает связывать пептиды с положительно заряженными аминокислотами, такими как аргинин (R) и лизин (K) на последней позиции связываемого кармана, а молекула HLA-A*01:01 связывается преимущественно с аспарагиновой кислотой (D) и глутаминовой кислотой (E) на третьей позиции пептида.

Таким образом, данные разделены не только по классам молекул, но и по аллелям, при этом объемы данных для разных аллелей различны.

## Построение интерпретируемой скрытой марковской модели с учетом особенностей задачи

Сформулируем требования к результату построения модели процессинга. Модель должна: учитывать различные длины последовательностей; позволять успешно находить связываемый участок пептида (выравнивать последовательности между собой); быть интерпретируемой.

Благодаря интерпретируемости возможен перенос знаний от модели для одного этапа процессинга на другой. Например, знания о том, какие аминокислоты связываемого участка участвуют в связывании молекулы, могут позволить исключить их из анализа иммуногенности пептида, так как связанные аминокислоты не вносят вклад в распознавание пептида напрямую [18].

Рассматриваемая задача предсказания представляет собой анализ и выравнивание связанных между собой последовательностей пептидов, в ходе которого осуществляется поиск структурных особенностей, описывающих рассматриваемый набор таких последовательностей. При этом под структурными особен-

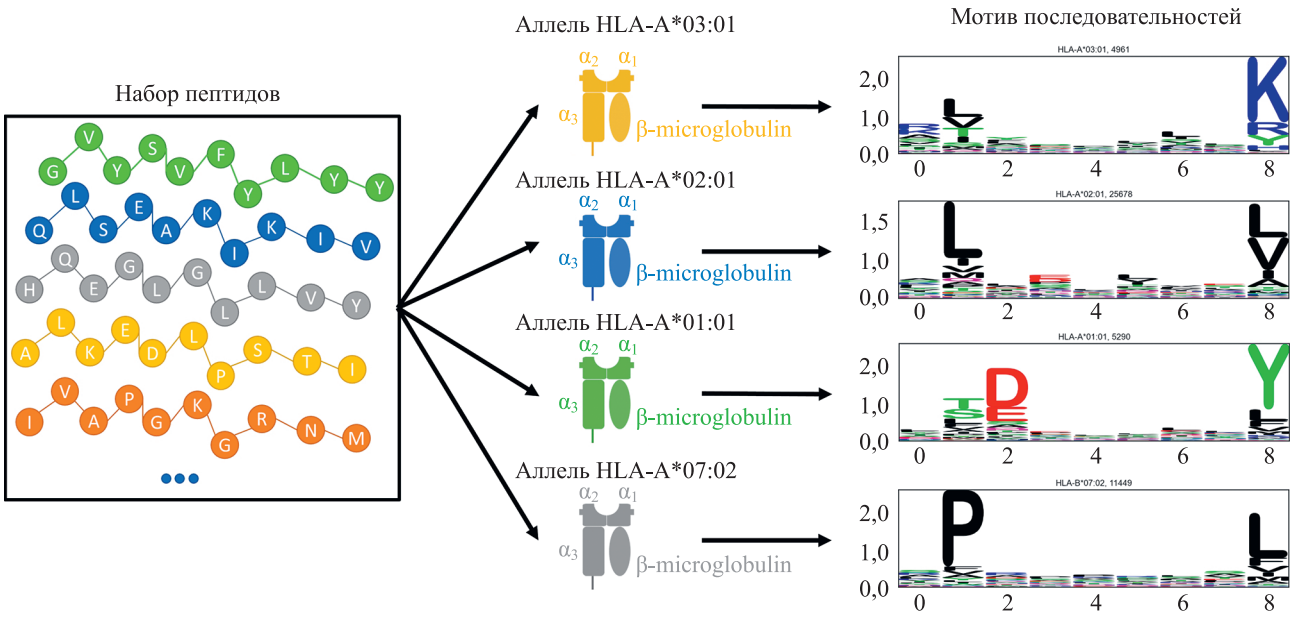



*Рис. 2*. Примеры мотивов связывания для нескольких аллелей

*Fig. 2*. Examples of Binding Motifs for various Alleles

ностями понимаются закономерности в расположении аминокислот внутри выравниваемого участка. Именно эти закономерности и являются информацией, извлекаемой из интерпретируемой модели.

Естественным способом представления информации о структурных особенностях последовательностей символов является граф выравнивания (рис. 3). Вершины графа соответствуют значениям аминокислот последовательностей, а ребра — возможным комбинациям аминокислот на позициях пептида. Кроме этого, структура содержит две дополнительные вершины, обозначающие начало и конец последовательности. Каждый отдельный путь внутри графа соответствует одной из последовательностей набора пептидов, а структура в целом получается объединением рассматриваемых выравненных последовательностей. В ходе объединения совпадающие аминокислоты разных пептидов, располагающиеся на той же позиции выравниваемого фрагмента, объединяются в общую вершину, «сшивая» между собой пути пептидов.

Любая модель, используемая для анализа набора последовательностей, соответствует графу выравнивания, восстановленному с определенной степенью точности. При этом чем точнее граф, тем лучше модель описывает обучающую выборку. Под интерпретируемостью модели понимается возможность анализировать полученный граф модели, извлекая данные о его ребрах, вершинах и фрагментах последовательностей анализируемых пептидов, соответствующих им. Предполагается, что каждая входная последовательность пептида будет выравнена вдоль одного из путей внутри результирующей модели. Каждый такой путь характеризует некоторые структурные особенности связываемого участка части пептидов обучающей выборки (свойственных данной молекуле ГКГС). Именно анализ таких путей позволит описать эти структурные особенности. Суть анализа в предложенном методе состоит в изучении характеристик узлов, входящих в рассматриваемый путь, а также набора пептидов, соответствующих ему.

Например, по рис. 3 можно сделать вывод, что в обучающей выборке (наборе пептидов, использованном для обучения модели) присутствуют пептиды двух типов (выделены цветом). Первый тип содержит в своем связывающем кармане (область между состояниями S5 и S11) соединение аминокислот S–S и встречается в длинных пептидах, второй тип участка (A–L) характерен для коротких пептидов.

Таким образом, задача построения модели заключается в построении графа выравнивания.

Скрытые марковские модели являются примером моделей, в которых граф, схожий с графом выравнивания, оптимизируется в явном виде [11]. В таких моделях задача анализа последовательностей пептидов может быть рассмотрена как задача анализа последовательностей наблюдений. Таким образом, каждая последовательность создана некоторым набором распределений (состояний) возможных исходов (например, символов). При этом наблюдение двух последовательных символов из различных распределений свидетельствует о связи между этими распределениями — существует вероятность перейти из одного состояния в другое при переходе к следующему символу.

Отличительная особенность полученных моделей — возможность учета различных длин входных последовательностей. Кроме того, результирующий граф состояний представляет собой выравнивание исходных последовательностей [13] и доступен для анализа в любой момент времени, что делает модель интерпретируемой. Комбинация этих качеств результирующей модели соответствует заявленным требованиям к методу построения и делает скрытые марковские модели основным кандидатом для анализа различных этапов процессинга пептидов.

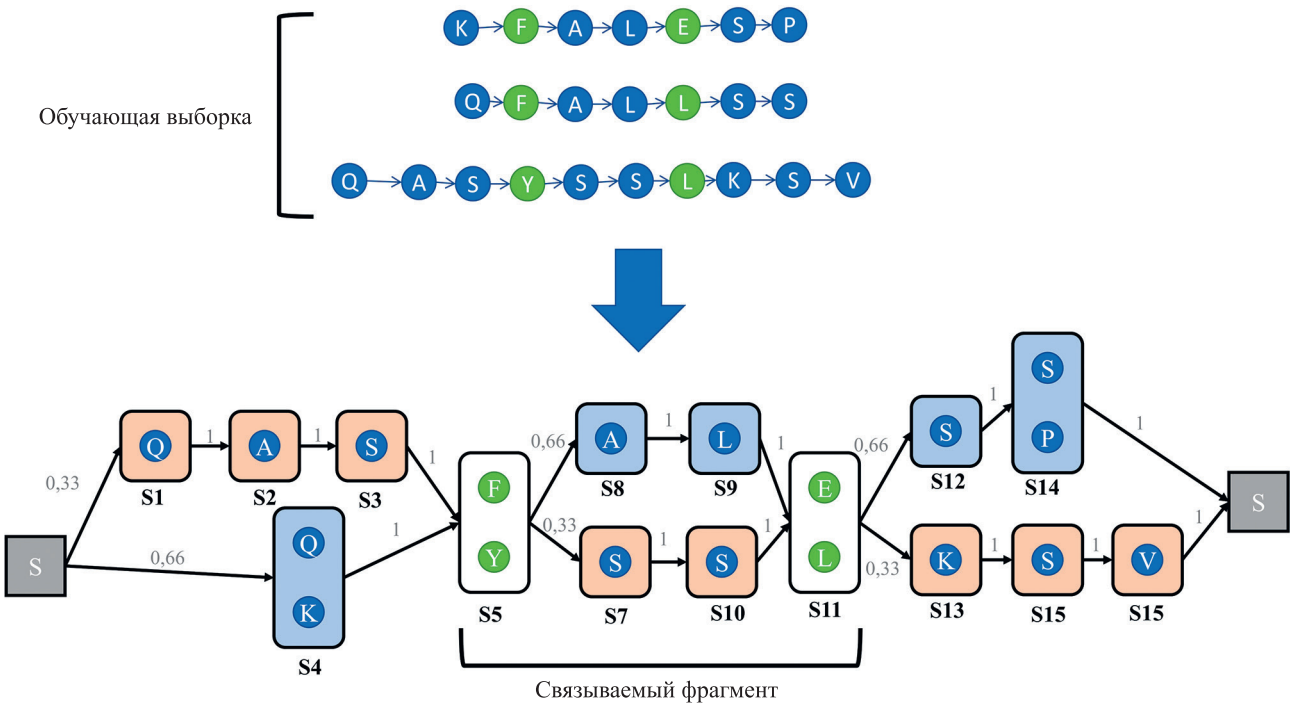



*Рис. 3*. Визуализация модели в виде графа выравнивания

*Fig. 3*. Visualizing models via alignment graphs

Для обучения целевой модели связывающей способности выберем алгоритм тренировки, ее параметры и архитектуру.

**Алгоритм тренировки моделей.** Процесс тренировки скрытых марковских моделей представляет собой итеративную процедуру обновления трех структур: матрицы переходов между состояниями, матрицы параметров распределений внутри состояний и вектора начальных вероятностей для состояний. Классическими и наиболее используемыми в настоящее время алгоритмами нахождения новых параметров являются алгоритмы Баума–Велша и Витерби [11, 23], которые являются частным случаем Expectation-Maximization (EM)-алгоритма [24]. В данных алгоритмах параметры распределений состояний модели обновляются итеративно в соответствии с символами (наблюдениями) последовательности из обучающей выборки. При этом конкретные распределения для обновления выбираются вдоль путей внутри графа модели, что является ключевым различием алгоритмов.

Алгоритм Баума–Велша использует процедуру прямого-обратного хода для поиска всех возможных путей выравнивания для текущей последовательности (с учетом текущих параметров распределений) и обновляет параметры модели вдоль найденных путей. Задача алгоритма Витерби [25] — поиск внутри модели самого вероятного пути для последовательности и обновление параметров только вдоль этого пути.

Именно различие в числе обновляемых путей помогает выбрать необходимый алгоритм, используемый в методе построения интерпретируемой модели. На рис. 4 изображен результат тренировки моделей для одного и того же набора пептидов указанными выше алгоритмами. Каждая модель изображена в виде графа состояний, в узлах которого дискретные распределения аминокислот визуализированы с помощью библиотеки анализа мотивов последовательности [26]. На рис. 4 буква соответствует аминокислоте, а ее размер — вероятности наблюдения аминокислоты в текущем состоянии.

В процессе тренировки модели алгоритмом Баума–Велша (рис. 4, *a*) происходит обновление сразу по всем возможным путям, поэтому разница между результирующими путями в модели уменьшается. Каждое отдельное состояние перестает быть уникальным и может повторяться. Эта схожесть между путями и состояниями мешает выполнить анализ путей и сделать вывод о различиях подгрупп пептидов внутри рассматриваемой аллели ГКГС.

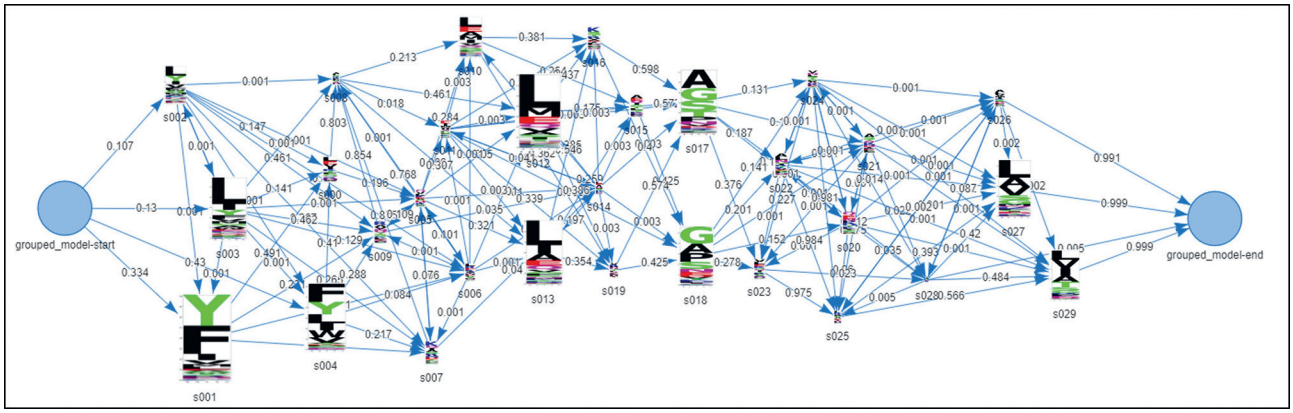
Тренировка с помощью алгоритма Витерби (рис. 4, *b*), благодаря обновлению единственного пути, делает каждое состояние более значимым и уникальным. Пути внутри модели сильно отличаются, что приводит к целесообразности их анализа.

Таким образом, в качестве алгоритма тренировки модели поиска и анализа связываемого участка целесообразно использовать алгоритм Витерби.

**Гиперпараметры тренировки.** Известна проблема использования алгоритма Витерби — склонность алгоритма сходиться к локальному минимуму [11]. Это делает состояния переобученными, и поэтому отдельные пути внутри модели являются неинформативными, так как им соответствует лишь небольшое число пептидов. Данная проблема может быть решена тренировкой ансамбля моделей и добавлением регуляризующих параметров, таких как:

— псевдонаблюдения — в процессе обновления распределения к каждому возможному наблюдению аминокислоты добавляется небольшое число на-

*a*




*b*

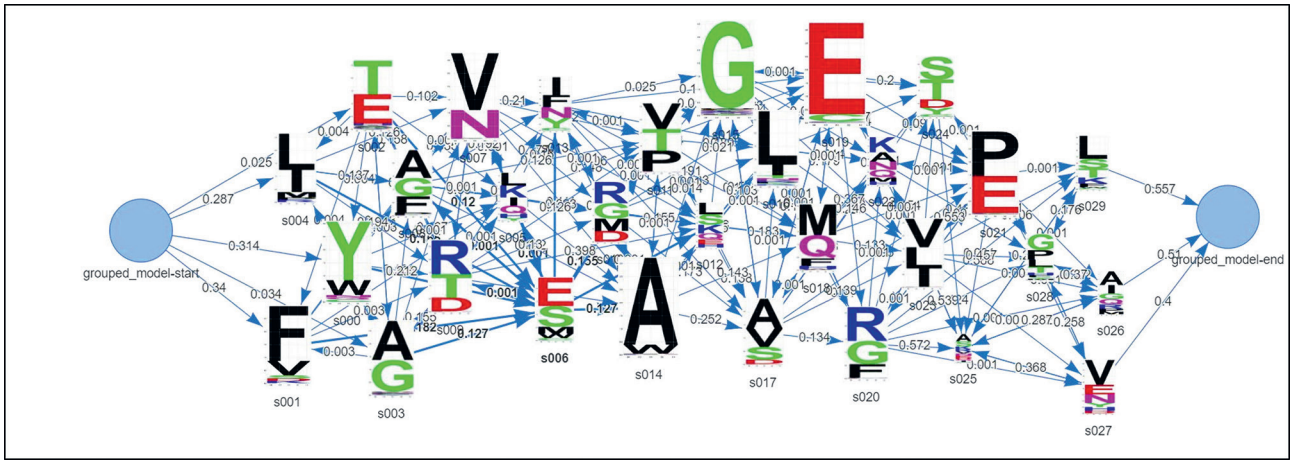



*Рис. 4*. Граф модели в результате обучения алгоритмами Баума–Велша (*a*) и Витерби (*b*)

*Fig. 4*. A model graph achieved through the Baum-Welch training algorithm (*a*) and the Viterbi training algorithm (*b*)

блюдений, что позволяет модели сохранить возможность перехода в текущее состояние и избежать его переобучения;

— инерция — коэффициент, с которым вероятности распределения на прошлой итерации алгоритма учитываются в процессе обновления распределения для текущей итерации. Он позволяет контролировать шаг обновления состояния.

**Архитектура модели.** Поиск и анализ связываемого участка в упрощенной форме представляют собой процедуру исследования марковского процесса [13], в котором последовательности наблюдений не содержат циклов. Каждый следующий символ соответствует новой аминокислоте в последовательности. При этом возможность возврата к предыдущим символам отсутствует. Данная особенность обусловлена линейностью последовательностей пептидов. Архитектура модели, поддерживающая такую структуру данных, должна представлять собой набор последовательно соединенных состояний или групп состояний. Именно такая архитектура выбрана и реализована в настоящей работе.

При этом можно выделить различные типы состояний внутри модели:

— состояния-якоря, которые представляют интерес в первую очередь, и предназначены для выравнивания якорей связываемого участка и соответствуют его началу. В таких состояниях будут искусственно поддерживаться большие значения вероятностей лишь для небольшого числа аминокислот, чтобы сделать состояния-якоря наиболее важными для всей модели;

— состояния-циклы обозначают неинформативную часть пептида и содержат распределения аминокислот, не относящихся к связываемому «карману»;

— обычные состояния, которые предназначены для выравнивания аминокислот на позициях, относящихся к связываемому участку пептида. Такие состояния объединены в последовательно соединенные группы и расположены между якорями.

Для того чтобы учесть структурные особенности классов молекул ГКГС, были предложены разные архитектуры для каждого из класса молекул (рис. 5). Для класса I использовано единственное состояние-цикл, переход к которому возможен только внутри выравниваемого фрагмента. Это обусловлено тем, что якоря для класса I располагаются сразу после начала пептида. Для класса II состояния-циклы размещены на краях модели, так как необходимо учитывать сдвиги связываемого фрагмента при выравнивании пептидов.

### Обучение и извлечение информации из моделей

Для обучения модели из базы данных связываемых пептидов [27] были выбраны последовательности од-

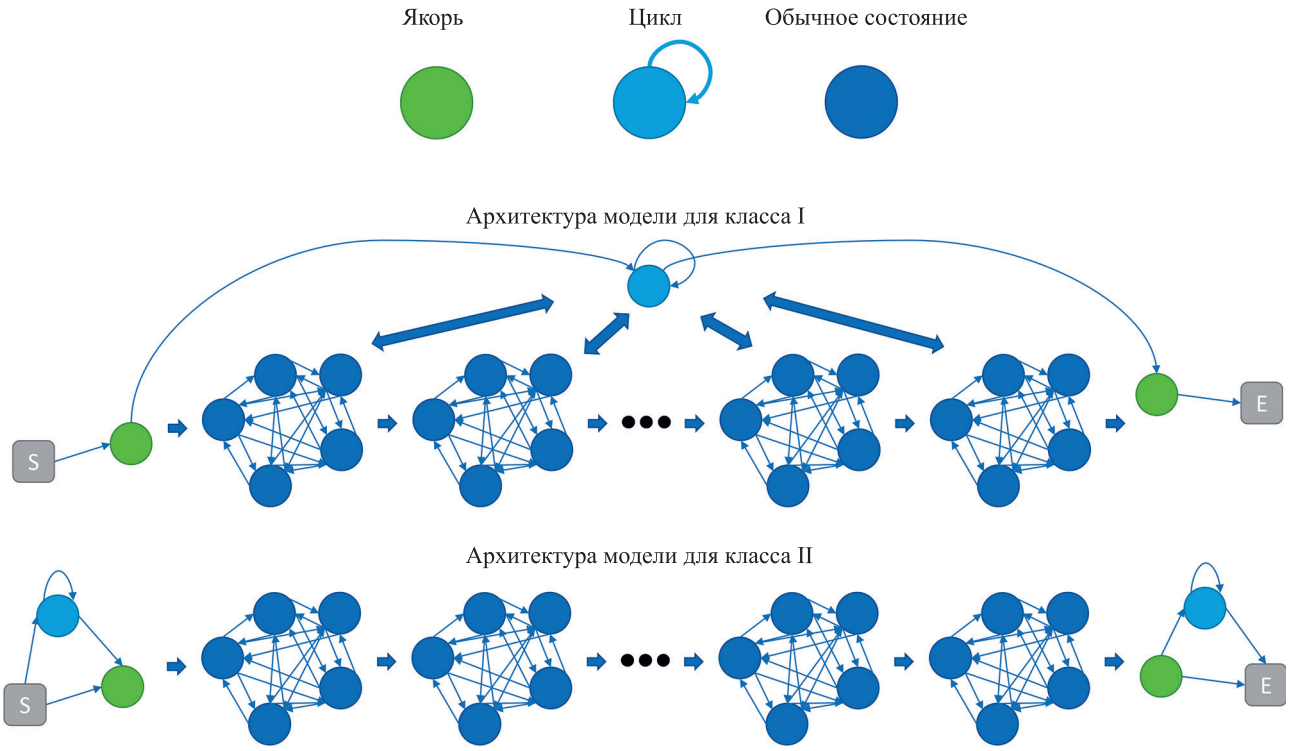



*Рис. 5*. Архитектура моделей
*Fig. 5*. Models architecture

ной аллели для каждого класса молекул ГКГС (аллель HLA-A*03-01 для класса I, аллель HLA-DRB1*01-01 для класса II). Фильтрация и объединение данных разных источников выполнены в соответствии с используемыми подходами [6, 15] с применением линейных пептидов и стандартных аминокислот.

В качестве экспериментальной архитектуры в настоящей работе для двух классов аллелей выбрана самая простая, где в каждой группе содержится единственное состояние. Таким образом, результирующая модель для класса II содержит два цикла, два якоря и семь последовательно соединенных состояний, а модель для класса I – один цикл, два якоря и семь последовательно соединенных состояний. Описанные архитектуры являются достаточно грубыми приближениями графа выравнивания, так как все пути внутри модели пересекаются в подмножестве вершин, соответствующих связываемому фрагменту пептида. Однако использование циклов в архитектуре модели помогает строить пути произвольной длины, в которых позиция этого фрагмента отличается. Анализ таких путей позволит найти позицию связывающего фрагмента.

**Обучен ансамбль моделей.** Так как моделям свойственно сходиться к локальному минимуму, то из 100 запусков обучения в ансамбль отобраны 25 моделей, показавших наиболее высокие значения суммарной вероятности для набора данных.

Процедура обнаружения связываемого участка использовала алгоритм Витерби [25] для поиска наилучшего пути внутри модели. При этом маркером связывающего участка обозначались аминокислоты, соответствующие всем состояниям, кроме состояний-циклов.

На рис. 6 изображен результат обучения двух моделей, а также примеры разметки связываемых участков пептидов для каждого пептида.

### Интерпретация результатов обучения

Обученные модели имеют единственно возможный фрагмент путей, соответствующий связываемому фрагменту. Поэтому граф моделей описывает только общий вид такого фрагмента пептидов без возможности анализа его вариаций. Проверка корректности графа проведена визуальным сравнением распределений состояний модели с известным мотивом из открытой базы данных Major Histocompatibility Complex (MHC) Motif Viewer [28]. Заметим, что распределения в состояниях моделей визуально соответствуют распределениям аминокислот для соответствующих позиций связываемого кармана, что позволяет сделать вывод об успешном обучении.

Дополнительная валидация моделей осуществлена в ходе двух экспериментов по анализу пептидов связывания для молекул класса II. Так как для пептидов класса II возможны различные сдвиги связываемого участка внутри пептида, выполнена проверка способности модели определения сдвигов. Успешным результатом валидации стало извлечение корректных последовательностей связываемого участка для рассматриваемого набора пептидов. Полученная валидация является еще одним шагом интерпретации результатов обучения.

В первом эксперименте использованы последовательности пептидов базы данных Immune Epitope

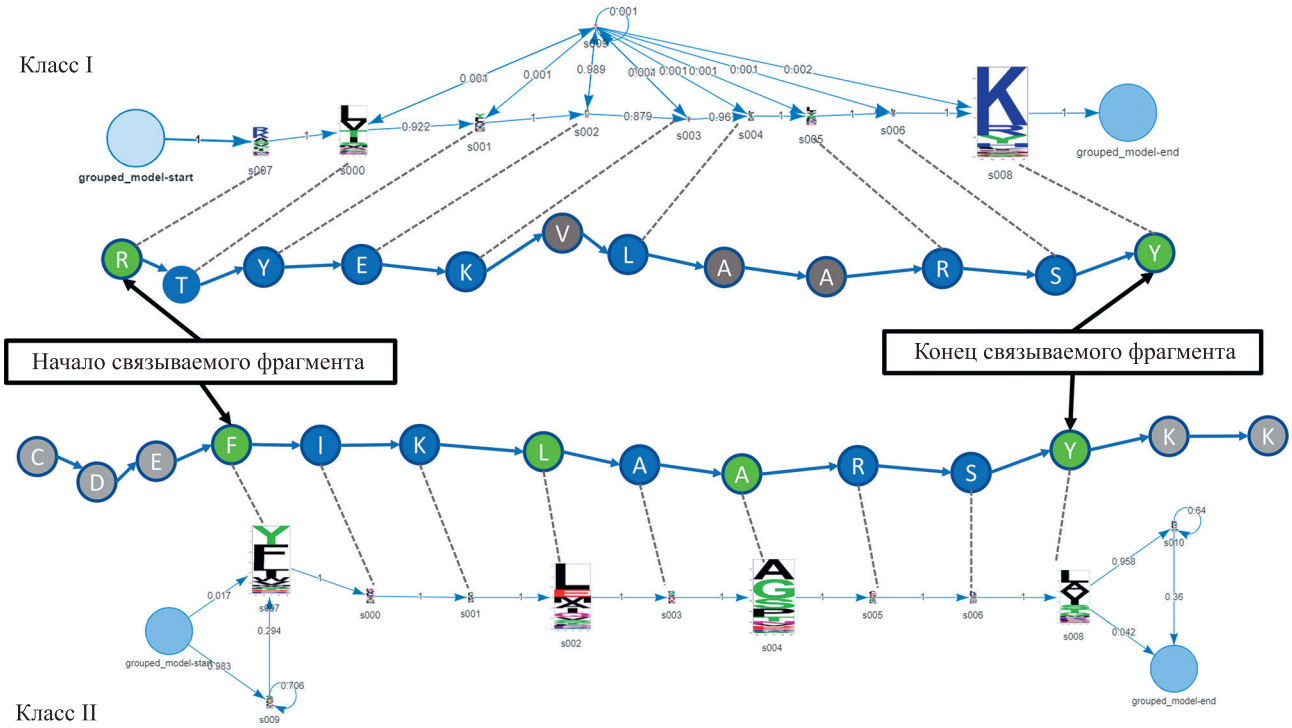



*Рис. 6*. Результат обучения и стратегия разметки связываемых участков для моделей I и II классов молекул

*Fig. 6*. Resultant models and a strategy for the identification of binding cores for the class I and class II models

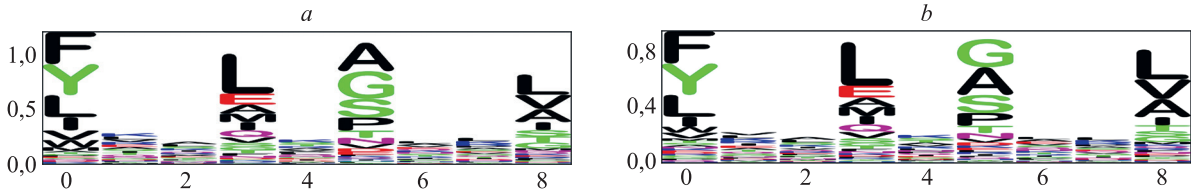



*Рис. 7*. Мотив связываемого участка из открытой базы данных Major Histocompatibility Complex Motif Viewer (*a*) и обнаруженный скрытой марковской моделью (*b*)

*Fig. 7*. Binding motif extracted from public database (*a*) and identified by the model (*b*)

Database [27]. Связывающие фрагменты последовательностей извлечены с применением обученных моделей. В результате визуальной валидации логотипов (рис. 7), подтверждено совпадение паттерна последовательностей исследованных фрагментов с искомым мотивом.

В ходе дополнительного эксперимента проверены связываемые фрагменты для последовательностей еще одного набора пептидов. Этот набор получен с помощью базы данных Protein Data Bank [29] и состоит из 14 кристаллографических белковых структур, для которых заранее известна связываемая последовательность. В результате эксперимента корректная последовательность связываемого фрагмента подтверждена для 13 из 14 рассмотренных структур.

## Выводы и обсуждение

Построенные модели (рис. 6) соответствуют заявленным требованиям к результату метода, включающим ограничения предметной области.

При этом открытым остается вопрос о возможности применения предложенного метода для построения моделей, которые можно было бы использовать для описания различных групп связываемых участков внутри последовательностей. Для этого необходимо натренировать модель на наборе пептидов, содержащем связывающие участки различных типов [30–32], и расширить этап анализа и интерпретации модели. Для тренировки предложено направление анализа путей внутри графа модели. При этом необходимо в дальнейшем увеличить число состояний внутри каждой группы модели.

Не менее важный вопрос следующих исследований: справится ли построенная модель с обнаружением позиции таких связываемых участков, которые повлекут неоднородность в данных.

Другим аспектом, достойным изучения, является применение иных видов распределений наблюдений внутри состояний графа модели. В работе использовано дискретное распределение аминокислот, где каждой соответствовала вероятность для наблюдения. Также интересным является вопрос о возможности использования для построения подобных моделей других характеристик пептидов, например, физико-химических свойств аминокислот [33].

## Заключение

В работе предложен метод построения, обучения и анализа интерпретируемых скрытых марковских моделей для задачи предсказания связывающей способности, а также произведено их обучение. Выбранные архитектуры и алгоритм обучения моделей позволили решить задачу обнаружения и визуализации главного паттерна связываемого участка, что показано в ходе двух экспериментов по анализу последовательностей пептидов.

## Литература


1. Chen D.S., Mellman I. Oncology meets immunology: The cancer-immunity cycle // Immunity. 2013. V. 39. N 1. P. 1–10. https://doi.org/10.1016/j.immuni.2013.07.012
2. Matsushita H., Vesely M.D., Koboldt D.C., Rickert C.G., Uppaluri R., Magrini V.J., Arthur C.D., White J.M., Chen Y.-S., Shea L.K., Hundal J., Wendl M.C., Demeter R., Wylie T., Allison J.P., Smyth M.J., Old L.J., Mardis E.R., Schreiber R.D. Cancer exome analysis reveals a T-cell-dependent mechanism of cancer immunoediting // Nature. 2012. V. 482. N 7385. P. 400–404. https://doi.org/10.1038/nature10755
3. Corradin G. Antigen processing and presentation // Immunology Letters. 1990. V. 25. N 1–3. P. 11–13. https://doi.org/10.1016/0165-2478(90)90082-2
4. Waldman A.D., Fritz J.M., Lenardo M.J. A guide to cancer immunotherapy: from T cell basic science to clinical practice // Nature Reviews Immunology. 2020. V. 20. N 11. P. 651–668. https://doi.org/10.1038/s41577-020-0306-5
5. Ott P.A., Hu Z., Keskin D.B., Shukla S.A. et al. An immunogenic personal neoantigen vaccine for patients with melanoma // Nature. 2017. V. 547. N 7662. P. 217–221. https://doi.org/10.1038/nature22991
6. Alspach E., Lussier D.M., Miceli A.P., Kizhvatov I., DuPage M., Luoma A.M., Meng W., Lichti C.F., Esaulova E., Vomund A.N., Runci D., Ward J.P., Gubin M.M., Medrano R.F.V., Arthur C.D., White J.M., Sheehan K.C.F., Chen A., Wucherpfennig K.W., Jacks T., Unanue E.R., Artyomov M.N., Schreiber R.D. MHC-II neoantigens shape tumour immunity and response to immunotherapy // Nature. 2019. V. 574. N 7780. P. 696–701. https://doi.org/10.1038/s41586-019-1671-8
7. Reynisson B., Alvarez B., Paul S., Peters B., Nielsen M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data // Nucleic Acids Research. 2020. V. 48. N W1. P. 449–454. https://doi.org/10.1093/nar/gkaa379
8. O'Donnell T.J., Rubinsteyn A., Laserson U. MHCflurry 2.0: Improved pan-allele prediction of MHC class I-presented peptides by incorporating antigen processing // Cell Systems. 2020. V. 11. N 1. P. 42–48. https://doi.org/10.1016/j.cels.2020.06.010
9. Phloyphisut P., Pornputtapong N., Sriswasdi S., Chuangsuwanich E. MHCSeqNet: a deep neural network model for universal MHC binding prediction // BMC Bioinformatics. 2019. V. 20. N 1. P. 270. https://doi.org/10.1186/s12859-019-2892-4
10. Shao X.M., Bhattacharya R., Huang J., Sivakumar I.K.A., Tokheim C., Zheng L., Hirsch D., Kaminow B., Omdahl A., Bonsack M., Riemer A.B., Velculescu V.E., Anagnostou V., Pagel K.A., Karchin R. High-throughput prediction of MHC class I and II neoantigens with MHCnuggets // Cancer Immunology Research. 2020. V. 8. N 3. P. 396–408. https://doi.org/10.1158/2326-6066.cir-19-0464
11. Rabiner L.R. A tutorial on hidden Markov models and selected applications in speech recognition // Proceedings of the IEEE. 1989. V. 77. N 2. P. 257–286. https://doi.org/10.1109/5.18626
12. Ревзин Л.М., Фильченков А.А., Тулупьев А.Л. Представление многозначных линейных по структуре скрытых марковских моделей в виде алгебраических байесовских сетей // Труды СПИИРАН. 2012. Т. 1. № 20. С. 186–199. https://doi.org/10.15622/sp.20.10
13. Eddy S.R. Profile hidden Markov models // Bioinformatics. 1998. V. 14. N 9. P. 755–763. https://doi.org/10.1093/bioinformatics/14.9.755
14. Bui H.-H., Sidney J., Peters B., Sathiamurthy M., Sinichi A., Purton K.-A., Mothé B.R., Chisari F.V., Watkins D.I., Sette A. Automated generation and evaluation of specific MHC binding


## References


1. Chen D.S., Mellman I. Oncology meets immunology: The cancer-immunity cycle. *Immunity*, 2013, vol. 39, no. 1, pp. 1–10. https://doi.org/10.1016/j.immuni.2013.07.012
2. Matsushita H., Vesely M.D., Koboldt D.C., Rickert C.G., Uppaluri R., Magrini V.J., Arthur C.D., White J.M., Chen Y.-S., Shea L.K., Hundal J., Wendl M.C., Demeter R., Wylie T., Allison J.P., Smyth M.J., Old L.J., Mardis E.R., Schreiber R.D. Cancer exome analysis reveals a T-cell-dependent mechanism of cancer immunoediting. *Nature*, 2012, vol. 482, no. 7385, pp. 400–404. https://doi.org/10.1038/nature10755
3. Corradin G. Antigen processing and presentation. *Immunology Letters*, 1990, vol. 25, no. 1–3, pp. 11–13. https://doi.org/10.1016/0165-2478(90)90082-2
4. Waldman A.D., Fritz J.M., Lenardo M.J. A guide to cancer immunotherapy: from T cell basic science to clinical practice. *Nature Reviews Immunology*, 2020, vol. 20, no. 11, pp. 651–668. https://doi.org/10.1038/s41577-020-0306-5
5. Ott P.A., Hu Z., Keskin D.B., Shukla S.A. et al. An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature*, 2017, vol. 547, no. 7662, pp. 217–221. https://doi.org/10.1038/nature22991
6. Alspach E., Lussier D.M., Miceli A.P., Kizhvatov I., DuPage M., Luoma A.M., Meng W., Lichti C.F., Esaulova E., Vomund A.N., Runci D., Ward J.P., Gubin M.M., Medrano R.F.V., Arthur C.D., White J.M., Sheehan K.C.F., Chen A., Wucherpfennig K.W., Jacks T., Unanue E.R., Artyomov M.N., Schreiber R.D. MHC-II neoantigens shape tumour immunity and response to immunotherapy. *Nature*, 2019, vol. 574, no. 7780, pp. 696–701. https://doi.org/10.1038/s41586-019-1671-8
7. Reynisson B., Alvarez B., Paul S., Peters B., Nielsen M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Research*, 2020, vol. 48, no. W1, pp. 449–454. https://doi.org/10.1093/nar/gkaa379
8. O'Donnell T.J., Rubinsteyn A., Laserson U. MHCflurry 2.0: Improved pan-allele prediction of MHC class I-presented peptides by incorporating antigen processing. *Cell Systems*, 2020, vol. 11, no. 1, pp. 42–48. https://doi.org/10.1016/j.cels.2020.06.010
9. Phloyphisut P., Pornputtapong N., Sriswasdi S., Chuangsuwanich E. MHCSeqNet: a deep neural network model for universal MHC binding prediction. *BMC Bioinformatics*, 2019, vol. 20, no. 1, pp. 270. https://doi.org/10.1186/s12859-019-2892-4
10. Shao X.M., Bhattacharya R., Huang J., Sivakumar I.K.A., Tokheim C., Zheng L., Hirsch D., Kaminow B., Omdahl A., Bonsack M., Riemer A.B., Velculescu V.E., Anagnostou V., Pagel K.A., Karchin R. High-throughput prediction of MHC class I and II neoantigens with MHCnuggets. *Cancer Immunology Research*, 2020, vol. 8, no. 3, pp. 396–408. https://doi.org/10.1158/2326-6066.cir-19-0464
11. Rabiner L.R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 1989, vol. 77, no. 2, pp. 257–286. https://doi.org/10.1109/5.18626
12. Revzin L.M., Filchenkov A.A., Tulupyev A.L. Representation of multinomial linear hidden Markov models in the form of algebraic Bayesian networks. *SPIIRAS Proceedings*, 2012, vol. 1, no. 20, pp. 186–199. (in Russian). https://doi.org/10.15622/sp.20.10
13. Eddy S.R. Profile hidden Markov models. *Bioinformatics*, 1998, vol. 14, no. 9, pp. 755–763. https://doi.org/10.1093/bioinformatics/14.9.755
14. Bui H.-H., Sidney J., Peters B., Sathiamurthy M., Sinichi A., Purton K.-A., Mothé B.R., Chisari F.V., Watkins D.I., Sette A. Automated generation and evaluation of specific MHC binding

predictive tools: ARB matrix applications // Immunogenetics. 2005. V. 57. N 5. P. 304–314. https://doi.org/10.1007/s00251-005-0798-y
15. Sarkizova S., Klaeger S., Le P.M., Li L.W., Oliveira G., Keshishian H., Hartigan C.R., Zhang W., Braun D.A., Ligon K.L., Bachireddy P., Zervantonakis I.K., Rosenbluth J.M., Ouspenskaia T., Law T., Justesen S., Stevens J., Lane W.J., Eisenhaure T., Zhang G.L., Clauser K.R., Hacohen N., Carr S.A., Wu C.J., Keskin D.B. A large peptidome dataset improves HLA class I epitope prediction across most of the human population // Nature Biotechnology. 2020. V. 38. N 2. P. 199–209. https://doi.org/10.1038/s41587-019-0322-9
16. Gomez-Perosanz M., Ras-Carmona A., Reche P.A. PCPS: A web server to predict proteasomal cleavage sites // Methods in Molecular Biology. 2020. V. 2131. P. 399–406. https://doi.org/10.1007/978-1-0716-0389-5_23
17. 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. Prediction of neo-epitope immunogenicity reveals TCR recognition determinants and provides insight into immunoediting // Cell Reports Medicine. 2021. V. 2. N 2. P. 100194. https://doi.org/10.1016/j.xcrm.2021.100194
18. Capietto A.H., Jhunjhunwala S., Pollock S.B., Lupardus P., Wong J., Hänsch L., Cevallos J., Chestnut Y., Fernandez A., Lounsbury N., Nozawa T., Singh M., Fan Z., de la Cruz C.C., Phung Q.T., Taraborrelli L., Haley B., Lill J.R., Mellman I., Bourgon R., Delamarre L. Mutation position is an important determinant for predicting cancer neoantigens // Journal of Experimental Medicine. 2020. V. 217. N 4. P. e20190179. https://doi.org/10.1084/jem.20190179
19. Andreatta M., Karosiene E., Rasmussen M., Stryhn A., Buus S., Nielsen M. Accurate pan-specific prediction of peptide-MHC class II binding affinity with improved binding core identification // Immunogenetics. 2015. V. 67. N 11–12. P. 641–650. https://doi.org/10.1007/s00251-015-0873-y
20. Punt J., Stranford S., Jones P., Owen J.A. Kuby Immunology. New York: Macmillan Education, 2019. 994 p.
21. Dendrou C.A., Petersen J., Rossjohn J., Fugger L. HLA variation and disease // Nature Reviews Immunology. 2018. V. 18. N 5. P. 325–339. https://doi.org/10.1038/nri.2017.143
22. Robinson J., Halliwell J.A., Hayhurst J.D., Flicek P., Parham P., Marsh S.G.E. The IPD and IMGT/HLA database: allele variant databases // Nucleic Acids Research. 2015. V. 43. N D1. P. D423–D431. https://doi.org/10.1093/nar/gku1161
23. Тулупьев А.Л., Николенко С.И., Сироткин А.В. Основы теории байесовских сетей. СПб.: Изд-во С.-Петерб. ун-та, 2019. P. 399.
24. Ng S.K., Krishnan T., McLachlan G.J. The EM algorithm // Handbook of Computational Statistics. 2012. P. 139–172. https://doi.org/10.1007/978-3-642-21551-3_6
25. Forney G.D. The viterbi algorithm // Proceedings of the IEEE. 1973. V. 61. N 3. P. 268–278. https://doi.org/10.1109/proc.1973.9030
26. Tareen A., Kinney J.B. Logomaker: beautiful sequence logos in Python // Bioinformatics. 2020. V. 36. N 7. P. 2272–2274. https://doi.org/10.1093/bioinformatics/btz921
27. Vita R., Mahajan S., Overton J.A., Dhanda S.K., Martini S., Cantrell J.R., Wheeler D.K., Sette A., Peters B. The immune epitope database (IEDB): 2018 update // Nucleic Acids Research. 2019. V. 47. N D1. P. D339–D343. https://doi.org/10.1093/nar/gky1006
28. Rapin N., Hoof I., Lund O., Nielsen M. MHC motif viewer // Immunogenetics. 2008. V. 60. N 12. P. 759–765. https://doi.org/10.1007/s00251-008-0330-2
29. Berman H.M. The protein data bank // Nucleic Acids Research. 2000. V. 28. N 1. P. 235–242. https://doi.org/10.1093/nar/28.1.235
30. Andreatta M., Lund O., Nielsen M. Simultaneous alignment and clustering of peptide data using a Gibbs sampling approach // Bioinformatics. 2013. V. 29. N 1. P. 8–14. https://doi.org/10.1093/bioinformatics/bts621
31. van Balen P., Kester M.G.D., de Klerk W., Crivello P., Arrieta-Bolaños E., de Ru A.H., Jedema I., Mohammed Y., Heemskerk M.H.M., Fleischhauer K., van Veelen P.A., Falkenburg J.H.F. Immunopeptidome analysis of HLA-DPB1 allelic variants reveals new functional hierarchies // The Journal of Immunology. 2020. V. 204. N 12. P. 3273–3282. https://doi.org/10.4049/jimmunol.2000192
32. Koşaloğlu-Yalçın Z., Sidney J., Chronister W., Peters B., Sette A. Comparison of HLA ligand elution data and binding predictions reveals varying prediction performance for the multiple motifs

predictive tools: ARB matrix applications. *Immunogenetics*, 2005, vol. 57, no. 5, pp. 304–314. https://doi.org/10.1007/s00251-005-0798-y
15. Sarkizova S., Klaeger S., Le P.M., Li L.W., Oliveira G., Keshishian H., Hartigan C.R., Zhang W., Braun D.A., Ligon K.L., Bachireddy P., Zervantonakis I.K., Rosenbluth J.M., Ouspenskaia T., Law T., Justesen S., Stevens J., Lane W.J., Eisenhaure T., Zhang G.L., Clauser K.R., Hacohen N., Carr S.A., Wu C.J., Keskin D.B. A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nature Biotechnology*, 2020, vol. 38, no. 2, pp. 199–209. https://doi.org/10.1038/s41587-019-0322-9
16. Gomez-Perosanz M., Ras-Carmona A., Reche P.A. PCPS: A web server to predict proteasomal cleavage sites. *Methods in Molecular Biology*, 2020, vol. 2131, pp. 399–406. https://doi.org/10.1007/978-1-0716-0389-5_23
17. 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. Prediction of neo-epitope immunogenicity reveals TCR recognition determinants and provides insight into immunoediting. *Cell Reports Medicine*, 2021, vol. 2, no. 2, pp. 100194. https://doi.org/10.1016/j.xcrm.2021.100194
18. Capietto A.H., Jhunjhunwala S., Pollock S.B., Lupardus P., Wong J., Hänsch L., Cevallos J., Chestnut Y., Fernandez A., Lounsbury N., Nozawa T., Singh M., Fan Z., de la Cruz C.C., Phung Q.T., Taraborrelli L., Haley B., Lill J.R., Mellman I., Bourgon R., Delamarre L. Mutation position is an important determinant for predicting cancer neoantigens. *Journal of Experimental Medicine*, 2020, vol. 217, no. 4, pp. e20190179. https://doi.org/10.1084/jem.20190179
19. Andreatta M., Karosiene E., Rasmussen M., Stryhn A., Buus S., Nielsen M. Accurate pan-specific prediction of peptide-MHC class II binding affinity with improved binding core identification. *Immunogenetics*, 2015, vol. 67, no. 11–12, pp. 641–650. https://doi.org/10.1007/s00251-015-0873-y
20. Punt J., Stranford S., Jones P., Owen J.A. *Kuby Immunology*. New York, Macmillan Education, 2019, 994 p.
21. Dendrou C.A., Petersen J., Rossjohn J., Fugger L. HLA variation and disease. *Nature Reviews Immunology*, 2018, vol. 18, no. 5, pp. 325–339. https://doi.org/10.1038/nri.2017.143
22. Robinson J., Halliwell J.A., Hayhurst J.D., Flicek P., Parham P., Marsh S.G.E. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Research*, 2015, vol. 43, no. D1, pp. D423–D431. https://doi.org/10.1093/nar/gku1161
23. Tulupev A.L., Nikolenko S.I., Sirotkin A.V. *Fundamentals of Bayesian Network Theory*. St. Petersburg, SPbU Publ., 2019, pp. 399. (in Russian)
24. Ng S.K., Krishnan T., McLachlan G.J. The EM algorithm. *Handbook of Computational Statistics*, 2012, pp. 139–172. https://doi.org/10.1007/978-3-642-21551-3_6
25. Forney G.D. The viterbi algorithm. *Proceedings of the IEEE*, 1973, vol. 61, no. 3, pp. 268–278. https://doi.org/10.1109/proc.1973.9030
26. Tareen A., Kinney J.B. Logomaker: beautiful sequence logos in Python. *Bioinformatics*, 2020, vol. 36, no. 7, pp. 2272–2274. https://doi.org/10.1093/bioinformatics/btz921
27. Vita R., Mahajan S., Overton J.A., Dhanda S.K., Martini S., Cantrell J.R., Wheeler D.K., Sette A., Peters B. The immune epitope database (IEDB): 2018 update. *Nucleic Acids Research*, 2019, vol. 47, no. D1, pp. D339–D343. https://doi.org/10.1093/nar/gky1006
28. Rapin N., Hoof I., Lund O., Nielsen M. MHC motif viewer. *Immunogenetics*, 2008, vol. 60, no. 12, pp. 759–765. https://doi.org/10.1007/s00251-008-0330-2
29. Berman H.M. The protein data bank. *Nucleic Acids Research*, 2000, vol. 28, no. 1, pp. 235–242. https://doi.org/10.1093/nar/28.1.235
30. Andreatta M., Lund O., Nielsen M. Simultaneous alignment and clustering of peptide data using a Gibbs sampling approach. *Bioinformatics*, 2013, vol. 29, no. 1, pp. 8–14. https://doi.org/10.1093/bioinformatics/bts621
31. van Balen P., Kester M.G.D., de Klerk W., Crivello P., Arrieta-Bolaños E., de Ru A.H., Jedema I., Mohammed Y., Heemskerk M.H.M., Fleischhauer K., van Veelen P.A., Falkenburg J.H.F. Immunopeptidome analysis of HLA-DPB1 allelic variants reveals new functional hierarchies. *The Journal of Immunology*, 2020, vol. 204, no. 12, pp. 3273–3282. https://doi.org/10.4049/jimmunol.2000192

recognized by HLA-DQ2.5 // Immunology. 2021. V. 162. N 2. P. 235–247. https://doi.org/10.1111/imm.13279
33. Kawashima S., Kanehisa M. AAindex: Amino Acid index database // Nucleic Acids Research. 2000. V. 28. N 1. P. 374–374. https://doi.org/10.1093/nar/28.1.374

32. Koşaloğlu-Yalçın Z., Sidney J., Chronister W., Peters B., Sette A. Comparison of HLA ligand elution data and binding predictions reveals varying prediction performance for the multiple motifs recognized by HLA-DQ2.5. *Immunology*, 2021, vol. 162, no. 2, pp. 235–247. https://doi.org/10.1111/imm.13279
33. Kawashima S., Kanehisa M. AAindex: Amino Acid index database. *Nucleic Acids Research*, 2000, vol. 28, no. 1, pp. 374–374. https://doi.org/10.1093/nar/28.1.374

## Авторы

**Клеверов Денис Анатольевич** — аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0009-0002-1362-486X, denklewer@gmail.com
**Шалыто Анатолий Абрамович** — доктор технических наук, профессор, профессор, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 56131789500, https://orcid.org/0000-0002-2723-2077, shalyto@mail.ifmo.ru
**Артемов Максим** — PhD, химические науки, профессор (исследователь), профессор, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация; профессор, Университет Вашингтона в Сент-Луисе. Медицинская Школа. Отдел патологии и иммунологии, Сент-Луис, 63110, США, SC 9242717500, https://orcid.org/0000-0002-1133-4212, martyomov@pathology.wustl.edu

*Статья поступила в редакцию 31.07.2023*
*Одобрена после рецензирования 29.08.2023*
*Принята к печати 30.09.2023*

## Authors

**Denis A. Kleverov** — PhD Student, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0009-0002-1362-486X, denklewer@gmail.com
**Anatoly A. Shalyto** — D.Sc., Full Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 56131789500, https://orcid.org/0000-0002-2723-2077, shalyto@mail.ifmo.ru

**Maxim N. Artyomov** — PhD (Chemistry), Professor (Researcher), ITMO University, Saint Petersburg, 197101, Russian Federation; Professor, Washington University in St. Louis. School of Medicine. Department of Pathology and Immunology, Saint Louis, 63110, USA, SC 9242717500, https://orcid.org/0000-0002-1133-4212, martyomov@pathology.wustl.edu

*Received 31.07.2023*
*Approved after reviewing 29.08.2023*
*Accepted 30.09.2023*