КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2020 Т. 12 № 6 С. 1383–1395
DOI: 10.20537/2076-7633-2020-12-6-1383-1395

Ки&М

АНАЛИЗ И МОДЕЛИРОВАНИЕ СЛОЖНЫХ ЖИВЫХ СИСТЕМ

УДК: 577.27

Применение ансамбля нейросетей и методов статистической механики для предсказания связывания пептида с главным комплексом гистосовместимости

И. В. Гребёнкин[1], А. Е. Алексеенко[2], Н. А. Гайворонский[1], М. Г. Игнатов[2], А. М. Казённов[2], Д. В. Козаков[3], А. П. Кулагин[1], Я. А. Холодов[1,a]

[1]Университет «Иннополис»,
Россия, 420500, г. Иннополис, ул. Университетская, д. 1
[2]Институт автоматизации проектирования РАН,
Россия, 123056, г. Москва, ул. 2-я Брестская, д. 19/18
[3]Университет Стони Брук,
США, г. Нью-Йорк, 11794, Stony Brook, 100 Nicolls Rd

E-mail: [a] kholodov@crec.mipt.ru

Получено 10.08.2020, после доработки — 19.10.2020.
Принято к публикации 29.10.2020.

Белки главного комплекса гистосовместимости (ГКГС) играют ключевую роль в работе адаптивной иммунной системы, и определение связывающихся с ними пептидов — важный шаг в разработке вакцин и понимании механизмов аутоиммунных заболеваний. На сегодняшний день существует ряд методов для предсказания связывания определенной аллели ГКГС с пептидом. Одним из лучших таких методов является NetMHCpan-4.0, основанный на ансамбле искусственных нейронных сетей. В данной работе представлена методология качественного улучшения архитектуры нейронной сети, лежащей в основе NetMHCpan-4.0. Предлагаемый метод использует технику построения ансамбля и добавляет в качестве входных данных оценку модели Поттса, взятой из статистической механики и являющейся обобщением модели Изинга. В общем случае модель отражает взаимодействие спинов в кристаллической решетке. Применительно к задаче белок-пептидного взаимодействия вместо спинов используются типы аминокислот, находящихся в кармане связывания. В предлагаемом методе модель Поттса используется для более всестороннего представления физической природы взаимодействия полипептидных цепей, входящих в состав комплекса. Для оценки взаимодействия комплекса «ГКГС + пептид» нами используется двумерная модель Поттса с 20 состояниями (соответствующими основным аминокислотам). Решая обратную задачу с использованием данных об экспериментально подтвержденных взаимодействующих парах, мы получаем значения параметров модели Поттса, которые затем применяем для оценки новой пары «ГКГС + пептид», и дополняем этим значением входные данные нейронной сети. Такой подход, в сочетании с техникой построения ансамбля, позволяет улучшить точность предсказания, по метрике положительной прогностической значимости (PPV), по сравнению с базовой моделью.

Ключевые слова: главный комплекс гистосовместимости, аффинность связывания, нейронная сеть, машинное обучение, модель Поттса

Исследование А. Е. Алексеенко и М. Г. Игнатова выполнено за счет гранта Российского научного фонда (проект № 19-74-00090). Исследование И. В. Гребёнкина выполнено за счет гранта Российского фонда фундаментальных исследований (проект № 19-37-90135 \19).

COMPUTER RESEARCH AND MODELING
2020 VOL. 12 NO. 6 P. 1383–1395
DOI: 10.20537/2076-7633-2020-12-6-1383-1395

Ки&М

ANALYSIS AND MODELING OF COMPLEX LIVING SYSTEMS

UDC: 577.27

Ensemble building and statistical mechanics methods for MHC-peptide binding prediction

I. V. Grebenkin[1], A. E. Alekseenko[2], N. A. Gaivoronskiy[1], M. G. Ignatov[2], A. M. Kazennov[2], D. V. Kozakov[3], A. P. Kulagin[1], Ya. A. Kholodov[1,a]

[1]Innopolis University,
1 Universitetskaya st., Innopolis, 420500, Russia
[2]Institute of Computer Aided Design of the Russian Academy of Sciences,
19/18 2 Brestskaya st., Moscow, 123056, Russia
[3]Stony Brook University,
100 Nicolls Rd, Stony Brook, NY, 11794, U.S.A.

E-mail: [a] kholodov@crec.mipt.ru

Received 10.08.2020, after completion — 19.10.2020.
Accepted for publication 29.10.2020.

The proteins of the Major Histocompatibility Complex (MHC) play a key role in the functioning of the adaptive immune system, and the identification of peptides that bind to them is an important step in the development of vaccines and understanding the mechanisms of autoimmune diseases. Today, there are a number of methods for predicting the binding of a particular MHC allele to a peptide. One of the best such methods is NetMHCpan-4.0, which is based on an ensemble of artificial neural networks. This paper presents a methodology for qualitatively improving the underlying neural network underlying NetMHCpan-4.0. The proposed method uses the ensemble construction technique and adds as input an estimate of the Potts model taken from static mechanics, which is a generalization of the Ising model. In the general case, the model reflects the interaction of spins in the crystal lattice. Within the framework of the proposed method, the model is used to better represent the physical nature of the interaction of proteins included in the complex. To assess the interaction of the MHC + peptide complex, we use a two-dimensional Potts model with 20 states (corresponding to basic amino acids). Solving the inverse problem using data on experimentally confirmed interacting pairs, we obtain the values of the parameters of the Potts model, which we then use to evaluate a new pair of MHC + peptide, and supplement this value with the input data of the neural network. This approach, combined with the ensemble construction technique, allows for improved prediction accuracy, in terms of the positive predictive value (PPV) metric, compared to the baseline model.

Keywords: major histocompatibility complex, binding affinity, neural network, machine learning, Potts model

Citation: *Computer Research and Modeling*, 2020, vol. 12, no. 6, pp. 1383–1395 (Russian).

The work of A. E. Alekseenko and M. G. Ignatov is supported by the Russian Science Foundation under grant 19-74-00090.
The work of I. V. Grebenkin is supported by the Russian Foundation for Basic Research under grant 19-37-90135 \19.

Введение

Главный комплекс гистосовместимости (ГКГС) — семейство мембранных белков, присутствующих в большинстве клеток позвоночных организмов. Их ключевыми особенностями являются полигенность (например, у человека присутствует шесть генов ГКГС-I) и полиморфность (на сегодняшний день известно несколько тысяч аллелей) [Lundegaard et al., 2007; Robinson et al., 2015]. Основной функцией белков ГКГС-I является представление на поверхности клетки пептидов, полученных в процессе расщепления протеасомой внутриклеточных белков. Связывание пептида с ГКГС-I делает возможным его распознавания Т-клетками и является необходимым условием его иммуногенности. Т-клетки способны распознавать пептиды, произведенные патогеном, но при этом игнорируют пептиды здоровой клетки. Когда Т-клетка распознает пептид, происходит образование тернарного комплекса Т-клетки, ГКГС-I и пептида — первый шаг иммунного ответа. Высокая вариативность генов ГКГС-I приводит к тому, что набор представляемых ими пептидов свой для каждого человека, что может вызывать различия в иммунном ответе на одни и те же антигены [Rucevic et al., 2020], а также является главной причиной отторжения донорских тканей при пересадке органов, так как вероятность полного совпадения наборов белков ГКГС у донора и реципиента крайне низка.

Определение набора пептидов, продуцируемых в больных клетках и способных связаться с конкретной аллелью ГКГС, дает возможность создавать персонализированные антиген-специфичные Т-клетки. Это может быть сделано путем инъекции набора иммуногенных пептидов в организм, что приводит к выработке самим организмом поколения Т-клеток, реагирующих на этот пептид. Подобный подход был использован несколькими рабочими группами [Carreno et al., 2015; Sahin et al., 2017; Ott et al., 2017] для иммунотерапии рака. В то же время было показано, что лишь 1 из 200 потенциальных пептидов имеет аффинность связывания с ГКГС выше, чем порог иммуногенности [Nielsen et al., 2007]. Таким образом, особую важность приобретает задача прогнозирования связывания пептида с молекулой ГКГС.

Входными данными для задачи численного предсказания аффинности пептида к белку ГКГС являются первичные последовательности ГКГС-I и пептида. Целевой переменной является экспериментально измеренное значение их аффинности друг к другу либо получаемая из него категориальная характеристика связывания. Так как лишь небольшая часть белка ГКГС взаимодействует с пептидом, распространен подход использования для анализа аффинности не всей последовательности белка, а только тех аминокислот, которые находятся в связывающей полости (англ. binding groove) белка и должны контактировать с пептидом для успешного образования комплекса [Nielsen et al., 2007]. Эта подпоследовательность называется псевдопоследовательностью ГКГС, и в данной работе мы используем для описания белков ГКГС псевдопоследовательности, полученные с использованием метода, описанного в работе [Nielsen, Andreatta, 2016].

На сегодняшний день существует множество подходов, позволяющих предсказывать связываемость с высокой точностью. Согласно последним исследованиям, подходы на основе нейронных сетей позволяют достичь точности 98 % истинно положительных результатов для более чем 90 % комплексов «ГКГС + пептид» в задаче качественного предсказания связывания [Nielsen, Andreatta, 2016; Jurtz et al., 2017]. На текущий момент лучшие решения, такие как NetMHCpan-4.0 [Jurtz et al., 2017], DeepLigand [Zeng, Gifford, 2019], MHCflurry [O'Donnell et al., 2018], используют нейронные сети. В данной работе предлагаются методологические улучшения алгоритма NetMHCpan-4.0.

Существующие методы

Существующие методы для предсказания связываемости ГКГС с пептидом могут быть разделены на две группы. Первая группа методов предполагает создание отдельной модели для каждой аллели ГКГС. Такие методы носят название аллель-специфичных методов. Вторая

группа методов, именуемых пан-специфичными, в свою очередь, предполагает создание единой модели сразу для всех интересующих аллелей ГКГС. Кроме того, методы подразделяются по типу предсказываемой переменной. Одна из характеристик связываемости ГКГС с пептидом — аффинность белок-пептидного связывания. Другой характеристикой является вероятность связывания белка с пептидом. При этом в существующих на сегодняшний день наборах экспериментальных данных может быть представлена как аффинность связи белков (измеряемая в нмоль), так и категориальная величина, показывающая факт связывания молекулы ГКГС с пептидом. Как следствие, существующие решения NetMHC, NetMHCpan, MHCflurry и DeepLigand [Jurtz et al., 2017; Zeng, Gifford, 2019; O'Donnell et al., 2018; Lundegaard et al., 2008] по предсказанию связываемости ГКГС с пептидом используют при обучении оба типа характеристик: аффинность связи в числовом выражении и факт образования комплекса как категориальную переменную. В частности, MHCflurry и NetMHC предсказывают аффинность связывания, в то время как модели NetMHCpan-4.0 и DeepLigand используют комбинацию аффинности связывания и категориальной переменной для улучшения точности предсказания. Также методы можно разделить по типу представления входных данных. Упомянутые решения NetMHC, NetMHCpan, MHCflurry и DeepLigand используют представление ГКГС и пептида в виде их первичной структуры, т. е. буквенных последовательностей, кодирующих аминокислоты. Однако каждый метод имеет свои особенности.

В данной работе предлагается подход к обучению пан-специфичной модели, не уступающей по качеству предсказания аффинности существующим популярным методам. Показано, что пан-специфичные методы имеют высокую производительность, обусловленную способностью методов предсказывать связываемость пептида с любой аллелью ГКГС [Hoof et al., 2009; Nielsen et al., 2007]. Несмотря на высокую точность и широкую применимость, пан-специфичные методы имеют и свои недостатки. Во-первых, такие методы предполагают тренировку архитектуры на большом объеме подготовленных экспериментальных данных, содержащих информацию о связываемости пептида с ГКГС [Kim et al., 2009]. Все вышеописанные методы используют для этого базу данных IEDB (Immune Epitope Database) [Vita et al., 2019]. Однако данные в IEDB представлены неравномерно относительно длины пептида. Более 73 % записей соответствуют пептидами длины 9. Один из способов решения этой проблемы: обучать модели на пептидах длины 9, а затем предсказывать связываемость ГКГС с псевдопоследовательностями длины 9, получаемыми из пептидов другой длины [Lundegaard et al., 2008]. Во-вторых, большинство моделей используют показатель аффинности связывания ГКГС с пептидом в качестве основной целевой переменной. Однако характеристика аффинности — это не единственный фактор, обусловливающий образование комплекса [Zeng, Gifford, 2019]. С биологической точки зрения такие факторы, как уровень экспрессии белка в клетке, мотивы пептидов, получаемых при расщеплении белка протеасомой, и эффективность внутриклеточного транспорта пептидов белками TAP, также играют важную роль в связывании пептида с ГКГС.

Одним из лучших пан-специфичных методов, согласно [MHC I Automated Server Benchmarks, 2019], на сегодняшний день является NetMHCpan-4.0. Первой особенностью метода является подход к представлению входных данных, предполагающий энкодинг молекулы ГКГС в виде псевдопоследовательности фиксированной длины [Nielsen, Andreatta, 2016]. В качестве подхода к энкодингу последовательности аминокислот используется матрица BLOSUM [Tong, 2013]. В то же время пептиды приводятся к единой длине, равной 9, путем вставок и удалений аминокислот. При этом NetMHCpan учитывает не только структуру пептида, но и информацию о соответствующей аллели ГКГС. На первом этапе происходит тренировка архитектуры на парах «ГКГС + пептид», в которых пептид представлен длиной 9. Этот подход проиллюстрирован на рис. 1. Далее осуществляется тренировка на парах, в которых пептиды имеют длину, отличающуюся от 9, но при этом происходит трансформация последних к длине 9. Для пептидов длины меньше 9 вставка вспомогательной подпоследовательности производится в каждую возможную

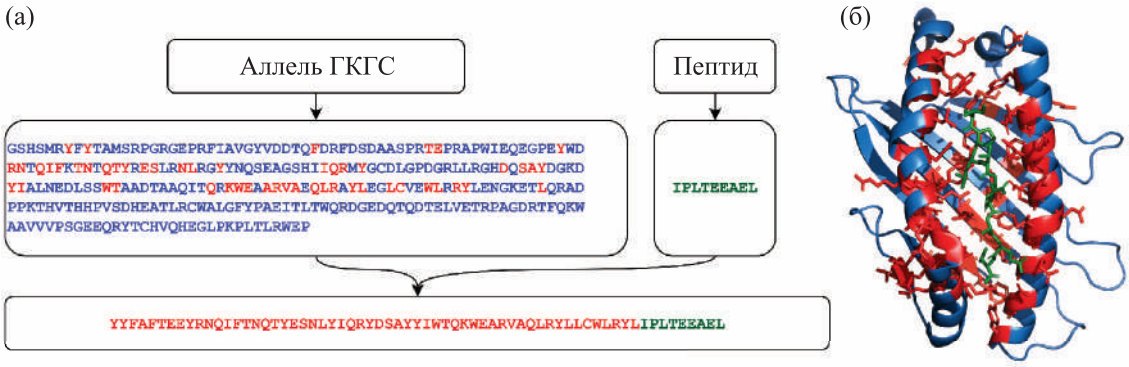

Рис. 1. (а) Диаграмма подготовки входных данных для нейронной сети на примера ГКГС HLA-B*35:01 и пептида IPLTEEAEL. Синими буквами обозначена последовательность белка, красными — выбранная согласно [Nielsen, Andreatta, 2016] псевдопоследовательность, зелеными — последовательность пептида. (б) Экспериментальная структура соответствующего комплекса (не использовалась при подготовке данных), PDB ID 5XOS [Shi et al., 2017]. Цвета структуры соответствуют цветам последовательности на диаграмме (а)

позицию, после чего полученные таким образом экземпляры подаются на вход претренированной модели. Таким образом, «удлиненный» пептид, на котором претренированная модель показывает лучший результат, отбирается в качестве лучшего и используется для предсказания значения аффинности. Для пептидов длины более 9 производятся удаления из всех возможных позиций. Полученные образцы аналогичным образом передаются на вход претренированной модели, что позволяет также отобрать наилучший из них. Помимо описанного механизма отбора наилучшей репрезентации комплекса, в котором пептиды отличной от 9 длины приводятся к длине 9, NetMHCpan использует дополнительные признаки с информацией о специфичности данного пептида по отношению к рассматриваемой аллели. Так, например, исходная длина пептида кодируется как категориальная переменная, также добавляется переменная, обозначающая длину последовательности, которая была вставлена/удалена.

NetMHCpan состоит из нескольких нейронных сетей и реализует технику сборки ансамбля. При этом используется перекрестная проверка с разбиением набора данных на 5 частей: каждая модель встречается в ансамбле 5 раз и тренируется на каждой части данных независимо. Каждая модель содержит 56 или 66 нейронов на скрытом слое и тренируется, используя 10 различных начальных конфигураций весов, сгенерированных случайным образом. В совокупности ансамбль NetMHCpan содержит 100 различных моделей.

В отличие от схожих методик [Zeng, Gifford, 2019; O'Donnell et al., 2018] подход NetMHCpan использует все сгенерированные модели для формирования ансамбля и не требует отдельного этапа селекции финальных моделей. Эта черта подхода NetMHCpan дает ему устойчивость к проблеме переобучения [Caruana et al., 2004].

Предлагаемые улучшения

Высокая производительность подхода NetMHCpan обусловлена несколькими составляющими: конфигурацией нейросетевой архитектуры, представлением входных данных (для длин пептидов отличных от 9) и техникой сборки в ансамбль. В качестве базовой архитектуры для обучения мы используем основной блок архитектуры NetMHCpan, но с небольшими изменениями в части репрезентации данных и выбора функции активации на выходном слое. По аналогии с архитектурой NetMHCpan в качестве входных данных используется вектор, состоящий из псевдопоследовательности ГКГС длины 34, сконкатенированной с последовательностью пептида.

Псевдопоследовательность ГКГС содержит только аминокислоты, лежащие в связывающей полости белка. Так как трехмерная структура белков ГКГС консервативна, выделение этих аминокислот из полной последовательности белка возможно методами гомологического анализа, как описано в работе [Nielsen, Andreatta, 2016]. Поскольку из исходного набора данных были отобраны только пептиды длины 9, в нашей работе размер входного слоя составил 43×1. Следующий слой — слой эмбеддинга, приводящий сконкатенированную последовательность к матрице размером 43×12. Перед передачей на полносвязный слой использован слой выравнивания, преобразующий матрицу в вектор длины 516. Полносвязный слой размерностью 70 нейронов на выходе имеет функцию активации ReLU, против Tanh, который используется в оригинальной архитектуре NetMHCPan. Функция активации ReLU выбрана для уменьшения проблемы затухания градиента при обратном распространении ошибки и тем самым повышения эффективности обучения. Выходной слой модели представлен одним нейроном с сигмоидальной функцией активации, дающей на выходе бинарное значение: связывается пептид с ГКГС или нет. Схема нейронной сети представлена на рис. 2. Начальные веса исходного слоя инициализировались нормальным распределением $\mathcal{N}(0, 1)$, для весов линейных слоев использовалась инициализация Ксавье (Xavier) [Glorot, Bengio, 2010]. Программный код написан на языке Python 3 с использованием библиотеки PyTorch [Paszke et al., 2019] для работы с нейронными сетями, а также вспомогательных библиотек NumPy [Van Der Walt et al., 2011], Pandas [McKinsey, 2010], Scikit-Learn [Pedregosa et al., 2011], Matplotlib [Hunter, 2007].

Таким образом, базовая архитектура схожа с оригинальной архитектурой NetMHCpan [Nielsen, Andreatta, 2016], но существенным дополнением, которое мы предлагаем, является использование оценки модели Поттса. Модель Поттса — это статистическая модель, описывающая вариативность некоторого заданного множества последовательностей аминокислот. Модель принимает на вход выровненные последовательности аминокислот (выравнивание позволяет лучшим образом представить схожие участки последовательностей) фиксированной длины K [Morcos et al., 2011]. Впоследствии, при получении на вход некоторой последовательности (псевдопоследовательности ГКГС, сконкатенированной с последовательностью пептида), модель Поттса дает оценку вероятности принадлежности входной последовательности к тому же классу, что и использованные при обучении последовательности. Базовая модель была расширена путем замены сигмоидальной функции активации последнего слоя на Tanh и добавлением еще одного полносвязного слоя, принимающего также оценку модели Поттса. К выходному значению, как и в базовой модели, применялась сигмоидальная функция для приведения результирующего значения к категориальному: 1 — в случае образования комплекса, 0 — в ином случае. Схема модифицированной нейронной сети представлена на рис. 3.

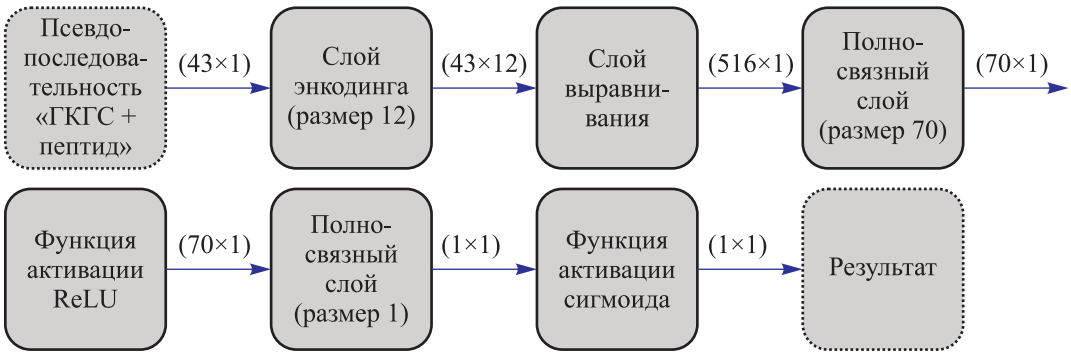

Рис. 2. Схема базовой архитектуры нейронной сети. Блоки изображают логические элементы схемы. Стрелки изображают переходы между логическими элементами, с указанием размерности данных на выходе блока

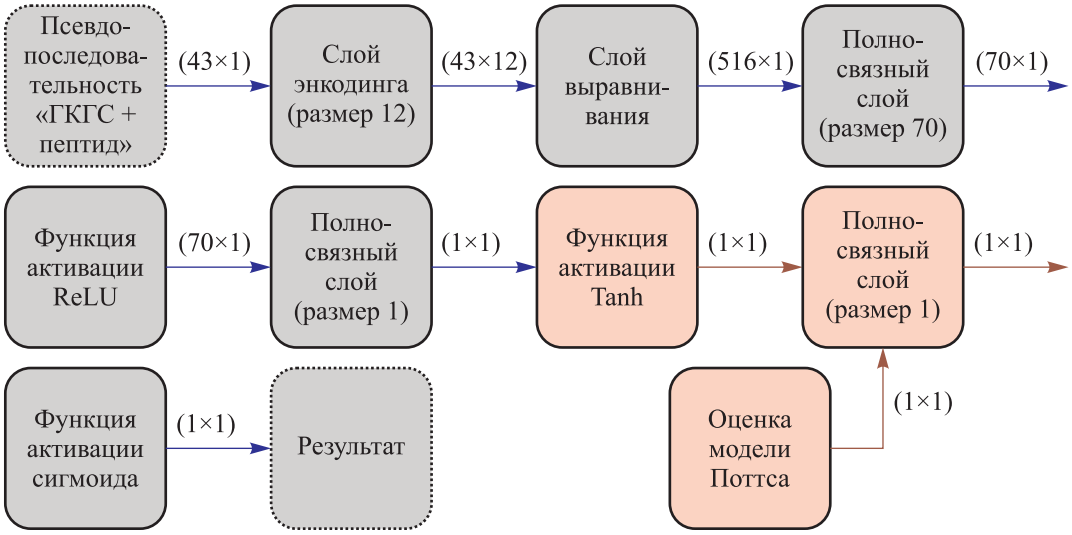

Рис. 3. Схема модифицированной архитектуры нейронной сети. Блоки изображают логические элементы схемы. Стрелки изображают переходы между логическими элементами, с указанием размерности данных на выходе блока. Оранжевым цветом выделены блоки, отсутствующие в базовой архитектуре

Параметризация модели Поттса осуществлялась с использованием библиотеки CCMpred [Seemayer et al., 2014]. При этом применялась видоизмененная оценочная функция:

$$P(a|J,h) = \frac{1}{Z}\exp\left(\sum_{i=1}^{L}\sum_{j=L+1}^{K} J_{ij}(a_i,a_j) + \sum_{i=1}^{K} h_i(a_i)\right),$$

где a — символ последовательности, J, h — параметры модели, Z — нормирующая константа, L — длина псевдопоследовательности ГКГС, K — суммарная длина ГКГС и пептида. Таким образом, учитывалось только взаимодействие между белком и пептидом. Значения оценочной функции Поттса подавались на вход расширенной архитектуре нейронной сети, производительность которой сравнивалась с базовой.

В процессе тренировки нейронной сети использовался оптимизатор Adam [Kingma, Ba, 2014]. Число тренировочных эпох, достаточное для достижения сходимости, составило 25. Размер обучающих мини-партий — 128. По аналогии с методологией NetMHCpan базовая модель и расширенная модель (оценка Поттса) тренировались с использованием техники сборки в ансамбль, то есть кооперативного обучения набора моделей и последующего объединения их результатов. Ансамблевые методы показали свою эффективность для решения большого спектра задач. Такой подход позволяет осуществить поиск наилучшей гипотезы в пространстве гипотез и на практике демонстрирует нивелирование проблем, связанных с переобучением модели на тренировочных данных [Chen, Shakhnarovich, 2014]. В представленной работе были обучены два ансамбля для базовой архитектуры и архитектуры с оценочной функцией Поттса. Размерность слоя эмбеддинга варьировалась от 12 до 21, что позволило обучить ансамбли, сформированные из 10 моделей.

Данные и метрики

Задача предсказания аффинности связывания ГКГС с пептидом представляет собой задачу бинарной классификации. Классическими метриками оценки качества моделей для данной задачи являются точность (precision), правильность (accuracy), ROC-кривая и положительная прогностическая значимость (PPV). В данной работе в качестве основной метрики мы используем

PPV. Это обусловлено тем, что при работе с наборами данных, в которых превалируют образцы одного класса, оценка качества с помощью ROC-кривой будет смещенной. Наборы данных комплексов «ГКГС–пептид» практически всегда имеют недостаток образцов с позитивным значением аффинности. Следовательно, метрика PPV представляется более чувствительной к ошибкам классификатора, чем ROC-кривая. Также PPV лучше соответствует целям применения полученного классификатора. Как правило, нам необходимо из набора аллелей ГКГС и пептидов найти пары, имеющие высокую аффинность связывания и, как следствие, вызывающие иммунный ответ, т. е. минимизировать число ложно-отрицательных предсказаний и повысить относительный ранг истинно-положительных. С другой стороны, ложно-положительные результаты относительно «безвредны», так как предсказания модели будут проверяться *in vitro*.

Расчет PPV производился следующим образом. Определяется значение n так, чтобы в наборе данных присутствовало по крайней мере n образцов, связывающихся в эксперименте, и 99 n несвязывающихся образцов. Далее, выбранные 100 n образцов упорядочиваются по убыванию предсказанной моделью вероятности связывания и рассчитывается количество истинно связывающихся образцов среди первых n. Таким образом, если модель демонстрирует более высокие результаты для истинно отрицательных образцов, рассчитанное низкое значение PPV позволяет учитывать это. Данная метрика лучше отражает практическое применение полученной системы: большинство пептидов, представленных в организме, не связываются с конкретной аллелью ГКГС, и среди них необходимо найти те, которые образуют комплекс.

Для тренировки моделей был использован оригинальный набор данных NetMHCpan-4.0, основанный на базе данных IEDB [Vita et al., 2019]. Была проведена дополнительная предобработка, включающая удаление дубликатов и нормализацию имен аллелей. Предлагаемая модель использует категориальную целевую переменную со значением 1 в случае образования комплекса и 0 в случае отсутствия связи ГКГС с пептидом. Часть экспериментальных данных в IEDB содержит численное значение аффинности. Для приведения его к категориальному виду каждая пара «ГКГС–пептид» считалась связывающейся, если ее аффинность была ниже 500 нмоль, и несвязывающейся иначе. Пороговое значение в 500 нмоль было выбрано, так как оно достаточно хорошо соответствует поведению Т-клеток [Sette et al., 1994].

Из общего набора данных был отобран набор, содержащий только пептиды длины 9. Таким образом, после этапа препроцессинга полный тренировочный набор содержит 3 634 949 негативных образцов и 134 835 позитивных. Набор, состоящий только из пептидов длины 9, содержит 541 398 негативных образцов и 78 992 позитивных.

Результаты

Для сравнения производительности моделей использовалась метрика положительной прогностической значимости (PPV). Так как различные аллели в разной степени представлены в IEDB, точность работы алгоритма на разных аллелях может значительно отличаться. Поэтому PPV рассчитывается для каждой аллели ГКГС из тестового набора отдельно.

Численный метод оценки предполагает сравнение отношения предсказаний пары методов $\frac{PPV_2(a)}{PPV_1(a)}$ с заранее выбранными пороговыми значениями. При использовании данного подхода были сформированы пять показателей, позволяющих достаточно хорошо оценить, насколько два метода отличаются по предсказательной способности. В случае если для аллели a значение $\frac{PPV_2(a)}{PPV_1(a)} < 0.7$ считается, что результат предсказания вторым методом на 30 % хуже. Для краткости принято обозначение «$< -30\,\%$». Если значение отношения превосходит порог $1.3\left(\frac{PPV_2(a)}{PPV_1(a)} \geq 1.3\right)$, считается, что второй метод демонстрирует результат на 30 % лучше по

сравнению с первым методом. Помимо порогового значения 30 %, также использовалось пороговое значение 10 %. Если соотношение результатов двух методов лежит в интервале 0.9–1.1, считается, что методы демонстрируют близкие значения производительности. На первом этапе нами была проверена производительность базовой архитектуры, которую предполагалось улучшать за счет интеграции оценки модели Поттса и техники сборки ансамбля. Для валидации производительности базовой архитектуры мы использовали сравнение достигнутого ей значения PPV с результатами, полученными с помощью NetMHCpan. Базовая архитектура продемонстрировала улучшение результата более чем на 30 % только для 7 аллелей, более чем на 10 % также для 7 аллелей. При этом результат, близкий к NetMHCpan, был получен для 28 аллелей. Ухудшение результата более чем на 10 % было получено для 16 аллелей, более чем на 30 % — для 14 аллелей. Примечательным является то, что для нескольких аллелей, для которых NetMHCpan показывает стабильно низкий результат (HLA-B:45*06, H2-Kd, HLA-B*14*02), базовая архитектура позволяет получить значительно более высокое значение. Можно заключить, что в среднем базовая архитектура демонстрирует сравнимый с полной версией NetMHCpan результат. Соответственно, архитектура представляется адекватной для последующих этапов, и на ее основе можно строить более сложные схемы.

Следующим этапом была проверка гипотезы об улучшении результатов за счет интеграции оценки модели Поттса в соответствии с приведенной выше схемой и использования техники сборки ансамбля. Была проведена серия экспериментов по тренировке ансамбля из 10 базовых моделей, тренировке архитектуры, расширенной добавлением оценки модели Поттса, и ансамбля из 10 моделей расширенной архитектуры. Результаты их сравнения с базовой моделью приведены на рис. 4.

Как видно из диаграммы, техника ансамблирования уже для 10 моделей базовой архитектуры позволяет значительно улучшить результат предсказания. Наличие небольшого количества аллелей, для которых результат ухудшился, — допустимый и прогнозируемый результат, который может быть нивелирован как подбором оптимального размера ансамбля, так и варьированием гиперпараметров на этапе тренировки модели. В то же время интеграция оценки модели Поттса в архитектуру базовой модели не дает видимого преимущества. С другой стороны, ансамбль

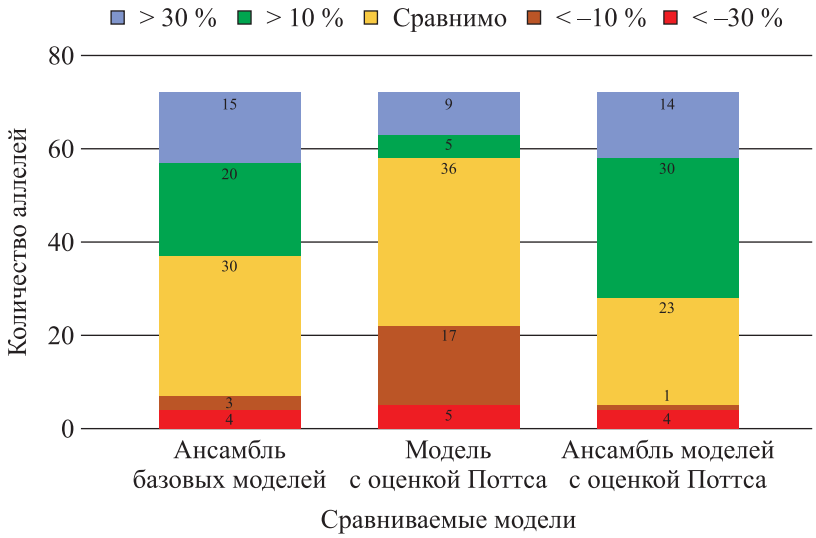

Рис. 4. Результаты сравнения модифицированных моделей с базовой. Каждый столбец графика изображает группы аллелей (и количество аллелей в группе), для которых произошло улучшение/ухудшение предсказательной способности при использовании модифицированной модели по сравнению с базовой, путем численного расчета оценки отношения PPV модифицированного метода к PPV базовой модели

10 моделей с оценкой Поттса демонстрирует лучший результат, чем ансамбль 10 базовых моделей. Можно заключить, что техника сборки ансамбля с дополненной оценкой модели Поттса архитектурой представляется методологически лучшим решением, имеющим перспективы для дальнейшего развития подхода.

Результаты сравнения ансамблей моделей с базовой архитектурой приведены на рис. 5 и 6, отображающих значения PPV (ось Y) для каждой аллели (ось X) в виде столбиковой диаграммы.

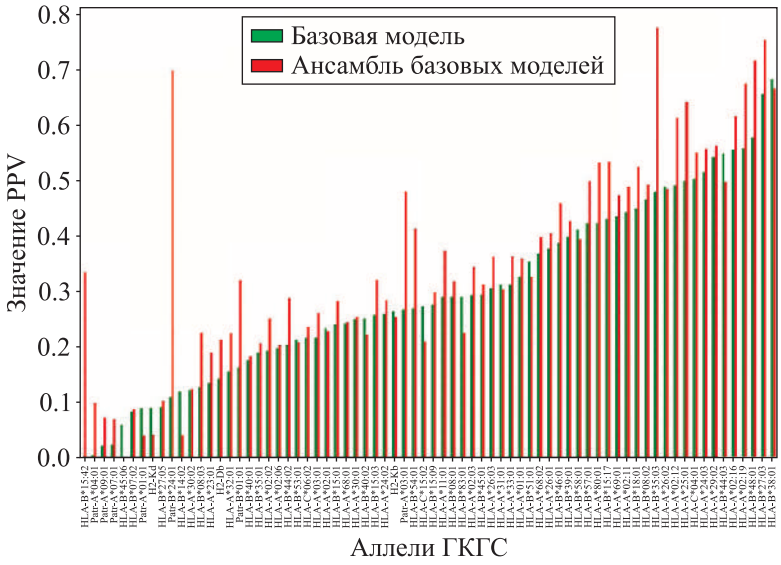

Рис. 5. Результаты сравнения производительности базовой модели с ансамблем базовых моделей. График демонстрирует значения PPV (положительная прогностическая значимость) для каждой аллели ГКГС из тестовой выборки, полученные базовой моделью (зеленый цвет) и ансамблем базовых моделей (красный цвет). Аллели отсортированы по значению PPV для лучшей интерпретируемости

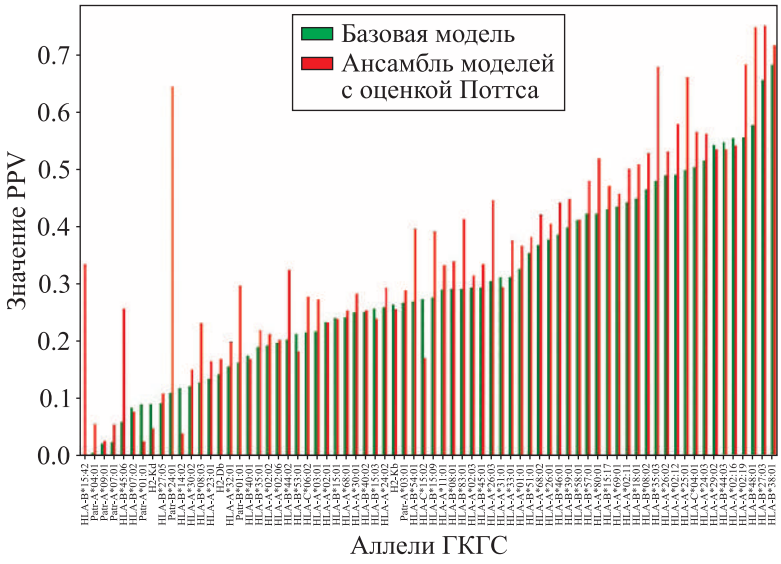

Рис. 6. Результаты сравнения производительности базовой модели с ансамблем моделей с оценкой Поттса. График демонстрирует значения PPV (положительная прогностическая значимость) для каждой аллели ГКГС из тестовой выборки, полученные базовой моделью (зеленый цвет) и ансамблем базовых моделей с оценкой Поттса (красный цвет). Аллели отсортированы по значению PPV для лучшей интерпретируемости

Заключение

В данной работе рассмотрена проблема предсказания аффинности белкового комплекса «ГКГС-I + пептид». Был рассмотрен подход NetMHCpan, на его основе была построена базовая модель, для которой предложены методологические улучшения путем встраивания оценки модели Поттса и использования техники ансамблирования. Была проведена серия экспериментов по тренировке базовой модели и модели с оценкой Поттса, а также проведены эксперименты с формированием ансамбля из данных архитектур. Предлагаемая в работе базовая архитектура продемонстрировала схожий с NetMHCpan-архитектурой результат по метрике PPV. Было показано, что техника сборки ансамбля улучшает результат предсказания аффинности белкового комплекса при использовании базовой архитектуры. В то же время предлагаемое методологическое улучшение посредством применения оценки модели Поттса в совокупности с формированием ансамбля моделей качественно улучшает получаемые результаты в сравнении с исходной архитектурой NetMHCpan при значительном снижении вычислительной сложности. Необходимо отметить, что подходы к нормализации длины пептида, использующиеся в NetMHCpan, применимы и к разработанной архитектуре, что дает ей потенциал для дальнейшего улучшения точности путем увеличения обучающей выборки и расширения области ее применения. Однако на сегодняшний день имеются альтернативные подходы к энкодингу последовательности аминокислот. Так, например, библиотека PyBioMed [Dong et al., 2018] предоставляет возможность производить для белкового комплекса расчет молекулярных дескрипторов, которые в свою очередь могут быть использованы для представления физико-химических характеристик белок-белковой связи. Другим возможным подходом является техника извлечения признаков из последовательности аминокислот [Asgari, Mohammad, 2015]. Данные подходы имеют потенциал для применения в рамках рассмотренной задачи и представляются интересными для последующих этапов исследования.

Список литературы (References)

Asgari E., Mohammad R. K. Continuous distributed representation of biological sequences for deep proteomics and genomics // PloS one. — 2015. — Vol. 10, No. 11.

Carreno B. M., Magrini V., Becker-Hapak M., Kaabinejadian S., Hundal J., Petti A. A. et al. A dendritic cell vaccine increases the breadth and diversity of melanoma neoantigen-specific T-cells // Science. — 2015. — Vol. 348, No. 6236. — P. 803–808.

Caruana R., Niculescu-Mizil A., Crew G., Ksikes A. Ensemble selection from libraries of models // Proceedings of the twenty-first international conference on Machine learning. — 2004. — P. 18.

Chen L., Shakhnarovich G. Learning ensembles of convolutional neural networks. — 2014.

Paszke A., Gross S., Massa F., Lerer A., Bradbury J., Chanan G. et al. PyBioMed: a python library for various molecular representations of chemicals, proteins and DNAs and their interactions // Journal of cheminformatics. — 2018. — Vol. 10, No. 1. — P. 16.

Glorot X., Bengio Y. Understanding the difficulty of training deep feedforward neural networks // Proceedings of the thirteenth International Conference on Artificial Intelligence and Statiscics. — 2010. — Vol. 9. — P. 249–256.

Henikoff S., Henikoff J. G. Amino acid substitution matrices from protein blocks // Proceedings of the National Academy of Sciences. — 1992. — Vol. 89, No. 22. — P. 10915–10919.

Hoof I., Peters B., Sidney J., Pedersen L. E., Sette A., Lund O., Buus S., Nielsen M. NetMHCpan, a method for MHC class I binding prediction beyond humans // Immunogenetics. — 2009. — Vol. 61, No. 1. — P. 1–13.

Hunter J. D. Matplotlib: A 2D graphics environment // Computing in Science & Engineering. — 2007. — Vol. 9, No. 3. — P. 90–95.

Jurtz V., Paul S., Andreatta M., Marcatili P., Peters B., Nielsen M. NetMHCpan-4.0: Improved peptide–MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data // The Journal of Immunology. — 2017. — Vol. 199, No. 9. — P. 3360–3368.

Kim Y., Sidney J., Pinilla C., Sette A., Peters B. Derivation of an amino acid similarity matrix for peptide: MHC binding and its application as a bayesian prior // BMC bioinformatics. — Vol. 10, No. 1. — P. 394.

Kingma D. P., Ba J. L. A method for stochastic optimization // arXiv:1412.6980. — 2014.

Lundegaard C., Lamberth K., Harndahl M., Buus S., Lund O., Nielsen M. NetMHC-3.0: Accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8–11 // Nucleic acids research. — 2008. — Vol. 36, No. 2. — P. 509–512.

Lundegaard C., Lund O., Keşmir C., Brunak S., Nielsen M. Modeling the adaptive immune system: Predictions and simulations // Bioinformatics. — 2007. — Vol. 23, No. 24. — P. 3265–3275.

MHC I Automated Server Benchmarks [Электронный ресурс]: http://tools.iedb.org/auto_bench/mhci/weekly/ (дата обращения: 01.06.2020).
MHC I Automated Server Benchmarks [Electronic resource]: http://tools.iedb.org/auto_bench/mhci/weekly/ (accessed: 01.06.2020).

Morcos F., Pagnani A., Lunt B., Bertolino A., Marks D. S., Sander C., Zecchina R., Onuchic J. N., Hwa T., Weigt M. Direct-coupling analysis of residue coevolution captures native contacts across many protein families // Proceedings of the National Academy of Sciences. — 2007. — Vol. 108, No. 49. — P. 1293–1301.

Nielsen M., Andreatta M. NetMHCpan-3.0: improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets // Genome medicine. — 2016. — Vol. 8, No. 1. — P. 33.

Nielsen M., Lundegaard C., Blicher T., Lamberth K., Harndahl M., Justesen S., Røder G., Peters B., Sette A., Lund O. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and-B locus protein of known sequence // PloS one. — 2007. — Vol. 2, No. 8.

Nielsen M., Lundegaard C., Worning P., Lauemøller S. L., Lamberth K., Buus S., Brunak S., Lund O. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations // Protein Science. — 2003. — Vol. 12, No. 5. — P. 1007–1017.

O'Donnell T. J., Rubinsteyn A., Bonsack M., Riemer A. B., Laserson U., Hammerbacher J. MHCflurry: Open-source class I MHC binding affinity prediction // Cell systems. — 2018. — Vol. 7, No. 1. — P. 129–132.

Ott P. A., Hu Z., Keskin D. B., Shukla S. A., Sun J., Bozym D. J. et al. An immunogenic personal neoantigen vaccine for patients with melanoma // Nature. — 2017. — Vol. 547. — P. 217–221.

Paszke A., Gross S., Massa F., Lerer A., Bradbury J., Chanan G. et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library // Advances in Neural Information Processing Systems. — 2019. — P. 8024–8035.

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O. et al. Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research. — 2011. — Vol. 12. — P. 2825–2830.

Peters M. E., Neumann M., Iyyer M., Gardner M., Clark C., Lee K., Zettlemoyer L. Deep contextualized word representations // arXiv preprint arXiv:1802.05365. — 2018.

Robinson J., Halliwell J. A., Hayhurst J. D., Flicek P., Parham P., Marsh S. G. The IPD and IMGT/HLA database: Allele variant databases // Nucleic acids research. — 2015. — Vol. 43, No. D1. — P. 423–431.

Rucevic M., Kourjian G., Boucau J., Blatnik R., Bertran W. G., Berberich M. J., Walker B. D., Riemer A. B., Le Gall S. Analysis of Major Histocompatibility Complex-Bound HIV Peptides Identified from Various Cell Types Reveals Common Nested Peptides and Novel T-Cell Responses // Journal of Virology. — 2016. — Vol. 90, No. 19. — P. 8605–8620.

Sahin U., Derhovanessian E., Miller M., Kloke B. P., Simon P., Löwer M. et al. Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer // Nature. — 2017. — Vol. 547. — P. 222–226.

Seemayer S., Gruber M., Soding J. CCMpred — fast and precise prediction of protein residue-residue contacts from correlated mutations // Bioinformatics. — 2014. — Vol. 30, No. 21. — P. 3128–3130.

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

Shi Y., Kawana-Tachikawa A., Gao F., Qi J., Liu C., Gao J., Cheng H., Ueno T., Iwamoto A., Gao G. F. Conserved Vδ1 binding geometry in a setting of locus-disparate pHLA recognition by $\delta/\alpha\beta$ T cell receptors (TCRs): insight into recognition of HIV peptides by TCRs // Journal of Virology. — 2017. — Vol. 91, No. 17. — P. e00725-17.

Tong J. C. BLOcks SUbstitution Matrix (BLOSUM). — Encyclopedia of Systems Biology. — NY: Springer, 2013.

Van Der Walt S., Colbert S C., Varoquaux G. The NumPy array: a structure for efficient numerical computation // Computing in Science & Engineering. — 2011. — Vol. 13, No. 2. — P. 22.

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. — P. 339–343.

McKinsey W. Data Structures for Statistical Computing in Python // Proceedings of the 9th Python in Science Conference. — 2010. — P. 56–61.

Zeng H., Gifford D. K. DeepLigand: Accurate prediction of MHC class I ligands using peptide embedding // Bioinformatics. — 2019. — Vol. 35, No. 14. — P. 278–283.