IN SILICO ОПРЕДЕЛЕНИЕ СПЕЦИФИЧНОСТИ НЕОАНТИГЕН-РЕАКТИВНЫХ Т-ЛИМФОЦИТОВ

***А.Е. Книга*[1,2]*, И.В. Поляков*[1,2], *А.В. Немухин*[1,2]**

[1]Московский государственный университет имени М.В. Ломоносова,
119991, Москва, Ленинские горы, 1, стр. 3; *эл. почта: kniga.ae@gmail.com
[2]Институт биохимической физики им. Н.М. Эмануэля РАН,
119334, Москва, ул. Косыгина, 4

Исследованы белок-белковые взаимодействия в системах TCR-pMHC на основе последовательностей, полученных в рамках глубокого секвенирования Т-лимфоцитов, окрашенных ДНК-мечеными тетрамерными комплексами неоантиген-MHC. Полноатомные модели вариабельных доменов иммунорецепторов неоантиген-реактивных Т-лимфоцитов в комплексе с пептидами, представленными молекулами MHC, были построены в рамках метода моделирования по гомологии. На основе полученной выборки были обучены классификаторы, с умеренной точностью способные различать аффинности тройных комплексов TCR-pMHC, что позволило проанализировать основные структурные детерминанты иммунного распознавания “чужого”. Мы предлагаем использовать *in silico* определение неоантигенной специфичности опухоль-инфильтрирующих Т-лимфоцитов для создания эффективных противораковых вакцин.

Ключевые слова: неоантиген; молекулярный докинг; белок-белковый комплекс; моделирование по гомологии; машинное обучение; TCR-pMHC

DOI: 10.18097/PBMC20216703251

ВВЕДЕНИЕ

Иммунотерапия рака получила признание наряду с традиционными хирургическими, радиационными и химиотерапевтическими методами терапии онкозаболеваний. Блокаторы контрольных точек, адоптивная терапия и неоантигенные вакцины представляют собой три её основных типа. Кроме того, развитие технологий секвенирования нового поколения повлекло за собой падение стоимости прочтения генома человека, что сделало доступным исследования геномных последовательностей клеток опухолей. В настоящее время развитие персонализированных методов лечения заболеваний требует обучения статистических моделей, способных быстро и эффективно интегрировать данные, полученные в результате прочтения ДНК, чтобы помочь принимать информированные терапевтически значимые решения [1]. Например, благодаря успехам масс-спектрометрии, удалось обучить модель нейронной сети NetMHCpan [2], которая позволяет с высокой точностью предсказывать энергию связывания произвольного пептида со специфическим аллелем MHC I и уже активно используется в клинических испытания неоантигенных вакцин.

Далеко не каждый неоантигенный пептид может вызывать иммунную активацию у цитотоксических лимфоцитов [3]. Для решения этой проблемы были разработаны модели, способные предсказывать иммуногенность отдельных неоантигенов в контексте всего иммунного репертуара. Альтернативным подходом может быть определение неоантигенной специфичности опухоль-инфильтрирующих лимфоцитов. Благодаря развитию методов анализа транскриптома, удаётся определить последовательности иммунорецепторов опухоль-инфильтрирующих Т-лимфоцитов [4]. Использование специализированных алгоритмов для поиска неоантигенов и анализ уровня их экспрессии позволяют получить список возможных вакцинных кандидатов [5]. Интегрируя эти два источника информации с помощью комбинации молекулярного моделирования и машинного обучения, можно осуществлять отбор тех неоантигенов, для которых вероятность амплификации иммунного ответа максимальна.

Т-лимфоциты осуществляют контроль над злокачественной трансформацией клеток, вызванной канцерогенами или онкогенными вирусами, с помощью распознавания пептидных фрагментов при участии белков главного комплекса гистосовместимости (MHC), представленных на поверхности клеточной мембраны. Каждый клон Т-лимфоцита несёт уникальный иммунорецептор (TCR), который обладает способностью распознавать новые пептиды одновременно с высокой чувствительностью и специфичностью. Для ответа на вопрос, каким образом достигается столь высокая эффективность распознавания, были предложены несколько механизмов [6].

1. Модель насыщения. Согласно данной модели, активация Т-лимфоцита зависит от числа тройных комплексов (ТК), образующихся на поверхности межклеточного контакта. Константа диссоциации тройного комплекса K_d определяет вероятность активации.

2. Модель кинетической коррекции. Необычайно высокая специфичность осуществляется за счёт контроля времени жизни тройного комплекса,

определяемого пороговым значением, необходимым для активации сигнального каскада. Небольшая разница аффинности отображается в значительное изменение времени жизни τ (и, соответственно, обратной ей по величине константе скорости диссоциации k_{off}) ценой увеличения времени существования межклеточного контакта. Такой подход не способен объяснить необычайно высокую чувствительность иммунного распознавания.

3. Кинетическая коррекция с модификациями. Приближение строится на основных посылках предыдущей модели, но дополнительно включаются факторы, направленные на объяснение оптимальной величины времени жизни TCR-pMHC, а также высокой чувствительности, например, за счёт активации одним и тем же pMHC нескольких TCR.

Таким образом, основными параметрами, характеризующими TCR-pMHC комплексы, являются k_{on}, k_{off} и K_d. Существует различие между значениями этих величин, измеренных в контексте свободной 3D диффузии лиганда перед связыванием с рецептором и в случае ограниченной мембраной 2D диффузии. Измерение 3D характеристик таких взаимодействий может осуществляться с помощью метода поверхностного плазмонного резонанса, либо с помощью клеточной сортировки. Измерения 2D параметров возможны, например, с помощью флуоресцентной микроскопии. Отличие 2D от 3D характеристик заключается в значительном увеличении k_{on} при исследовании в контексте 2D диффузии за счёт ограничений взаимной ориентации рецептора или лиганда на клеточных мембранах взаимодействующих клеток или в системах с использованием искусственной модели клеточной мембраны [6]. Тогда изменение константы k_{off} определяется воздействием тангенциально приложенной силы, вызванной локальной деформацией мембраны.

Несмотря на наличие в базах данных [7-9] значительного количества информации о 3D термодинамических характеристиках, единственными высокопроизводительными методами для получения информации о специфичности TCR-pMHC взаимодействия являются различные варианты клеточного сортинга [10, 11].

В настоящее время активно развиваются вычислительные методы предсказания антигенной специфичности. В основном они строятся либо на неявном представлении структур взаимодействующих белков-партнёров [12], либо с применением моделирования по гомологии, основанном на полноразмерных структурах тройных комплексов [11]. Значительное развитие получили вычислительные методы предсказания структур белков и белковых комплексов, а также различные методы высокопроизводительного докинга. Таким образом, эффективный вычислительный скрининг TCR-pMHC комплексов на основе структурных предсказаний является перспективным направлением молекулярной медицины.

МЕТОДИКА

Из литературных данных были получены аллели TRAV и TRBV иммунорецепторов неоантиген-реактивных Т-лимфоцитов [10], включая информацию об их антигенной специфичности. Информация об аллелях TRAJ и TRBJ, необходимая для восстановления полной аминокислотной последовательности TCR, но явным образом отсутствующая в дополнительных материалах, была получена с помощью поиска программой IgBlast [13] для приведённых данных о кодирующих участках CDR3αβ против набора последовательностей человеческих аллелей TRAJ и TRBJ. На основании данных об аллелях TRAV, TRBV и последовательностях CDR3, а также полученных данных о TRAJ и TRBJ, были восстановлены кодирующие последовательности α- и β-цепей Т-клеточных рецепторов с помощью программы stiTChR (https://github.com/JamieHeather/stitchr). Эти результаты были использованы для построения атомарных моделей с помощью процедуры моделирования по гомологии в приложении TCRmodel [14, 15], включающей поиск структурных шаблонов для консервативного и вариабельных фрагментов, их сшивание и минимизацию энергии по протоколу FastRelax [16, 17]. Для учёта пространственной подвижности CDR3 цепей был проведён конформационный поиск по методу NextGenerationKIC [18], ограниченный опцией “уточнение”, и для каждого рецептора были сгенерированы 10 моделей. По вышеописанному протоколу были построены модели 225 TCR.

Для моделирования конформаций неоантигенных пептидов в бороздке главного комплекса гистосовместимости также использовали подход на основе гомологии. Был произведён поиск на основе базы данных ATLAS. Для каждого пептида минимизировали расстояние Левенштейна, затем несколько структур (от 1 до 10) использовали для моделирования по гомологии с помощью протокола RosettaMHC [19], полученные структуры рассчитывали методом Монте-Карло в соответствии с протоколом FastRelax. Структуры с наименьшим значением оценочной функции REF2015 [20] использовали для последующих стадий расчёта. В рамках описанного протокола были построены модели 129 комплексов pMHC.

Определение ориентации взаимодействующих цепей иммунорецептора и комплекса неоантигенного эпитопа в бороздке главного комплекса гистосовместимости осуществляли на основе комплементарности поверхностей при помощи программного пакета молекулярного докинга PatchDock [21]. Таким образом, были получены параметры линейных преобразований (3 вращательных угла и 3 трансляционных параметра) координат молекулы лиганда в составе комплекса ТК TCR-pMHC. Для каждого линейного преобразования были рассчитаны: общая геометрическая комплементарность интерфейса (score), площадь образуемого интерфейса (Area), а также комплементарность с учётом лишь только участков эпитопа (as1),

паратопа (as2), или обоих участков (as12), а также атомарной энергии десольватации (ACE). Для ограничения конформационного пространства и ускорения процедуры молекулярного докинга использовали следующие геометрические критерии взаимного расположения молекул в комплексе TCR-pMHC: расстояния между центрами масс α- и β-цепей иммунорецептора и соответствующими им спиралями α2 и α1 главного комплекса гистосовместимости не должны превышать 3 нм; расстояния между центрами масс CDR3αβ и центром массы пептидного эпитопа не должны превышать 1,5 нм. Расчёт комплементарности взаимодействующих участков молекулярной поверхности был ограничен участками вариабельных петель CDR3αβ и последовательностью эпитопа для ускорения процедуры докинга. Среди полученных после процедуры докинга структур отбирали те, которые обладали наибольшей степенью геометрической комплементарности [21], нормированной на площадь образуемого интерфейса. Для отобранных структур производили оптимизацию геометрических параметров по методу Монте-Карло FastRelax с использованием параметров скрипта InterfaceRelax2019.

Для проверки статистических гипотез использовали парный t-тест. Для проверки нормальности распределения случайных величин использовали критерий Лиллиефорса с уровнем значимости α=0,05.

Обучение логистических моделей проводили с использованием L2-регуляризации с параметром C=1. Класс 1 (агонисты) присваивали тем ТК, для которых информация о связывании была получена из эксперимента по клеточной сортировке окрашенных тетрамерными ДНК-меченными pMHC Т-лимфоцитов. Для остальных ТК — 0 (не-агонисты). Так как выборки были несбалансированные, то использовали весовые коэффициенты классов в целевой функции кросс-энтропии, пропорциональные их представленности. Во всех случаях отбор дескрипторов для построения модели логистической регрессии осуществляли с применением дисперсионного анализа (ANOVA), уровень значимости α=0,05. Значения дескрипторов были нормированы на их среднее и стандартное отклонение.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Проверка способности оценочной функции Rosetta различать энергетические характеристики специфических лигандов соответствующих иммунорецепторов

Для проверки гипотезы о том, существуют ли какие-либо характерные отличия между структурными характеристиками ТК агонистов и не-агонистов, было проведено сравнение их энергетических и геометрических характеристик, рассчитанных для построенных по гомологии полноатомных моделей тройных комплексов двух классов: высокоаффинных (далее в тексте — агонисты) и низкоаффинных (далее — не-агонисты). Для статистической проверки гипотезы о равенстве средних значений изменения энергии при диссоциации тройных комплексов специфического (агонист) и неспецифического (не-агонист) классов были построены полноатомные модели. Было рассмотрено 200 иммунорецепторов, содержащих специфические (по 5 конфигураций лиганда, полученных из линейных преобразований с наибольшим значением геометрической комплементарности для каждого иммунорецептора в комплексе с его нативным антигенным пептидом) и неспецифические лиганды (также 5 конфигураций на каждый иммунорецептор, но для 5 разных случайно выбранных пептидов не-агонистов). На этом этапе были оптимизированы структуры 2000 моделей ТК. Среди полученных комплексов для каждого класса (агонист/не-агонист) отбирали те модели, которые обладали наибольшим абсолютным значением рассчитанной разницы нормированной энергии при диссоциации TCR-pMHC в группе моделей одного ТК (из 10 моделей на 1 ТК, отбирали 2 модели: одна модель ТК с агонистом, вторая для ТК с не-агонистом). Таким образом, для каждого из 200 иммунорецепторов измеряли изменение энергетических характеристик образованного данным иммунорецептором ТК при замене его пептида с агониста на не-агонист.

Для статистической проверки гипотезы о равенстве средних разниц энергии диссоциации тройных комплексов специфического (агонист) и неспецифического (не-агонист) типа был произведён расчёт разниц энергии диссоциации (измеряется в единицах энергии программного пакета Rosetta “Rosetta Energy Unit”, далее REU) и изменений площади поверхности доступной растворителю (ΔSASA). Полученные значения нормированных изменений энергии диссоциации ΔG/ΔSASA (1) двух выборок были использованы для проведения парного t-теста.

$$\Delta G_i=(E_{TCRi}+E_{pMHC}-E_{TCRi-pMHC})/\Delta SASA,\ [REU/nm^2] \quad (1);$$

$$\Delta\Delta G_i = \Delta G^{B} - \Delta G^{NB},\ \text{для } TCR_i,\ [REU/nm^2] \quad (2).$$

Нулевая гипотеза (**H_0**) принимается, если нет разницы, в среднем, между изменением энергии диссоциации для структур ТК с агонистом — высокоаффинным $pMHC^{B}$ (ΔG^{B}) и не-агонистом — случайным антигеном $pMHC^{NB}$ (ΔG^{NB}) (2), то есть $\mu(\Delta\Delta G_i)=0$. Альтернативная гипотеза (**H_1**) принимается, если есть отрицательная разница, в среднем, между изменением энергии диссоциации для структур ТК с агонистом и не-агонистом, то есть $\mu(\Delta\Delta G_i)<0$. Было получено значение парного одностороннего t-критерия $t(200)=-3{,}83$ и $p=1{,}67\times10^{-4}$, следовательно, нулевую гипотезу **H_0** можно отвергнуть.

Для статистической проверки гипотезы о равенстве средних изменений энергии диссоциации тройных комплексов специфического и неспецифического связывания при замене всех неякорных остатков антигена на аланин (pA) использовали модели предыдущего эксперимента (2000 ТК WT) для моделирования по методу Монте-Карло FastDesign. В результате были получены дополнительные 2000 структур ТК pA для 200 комплексов TCR

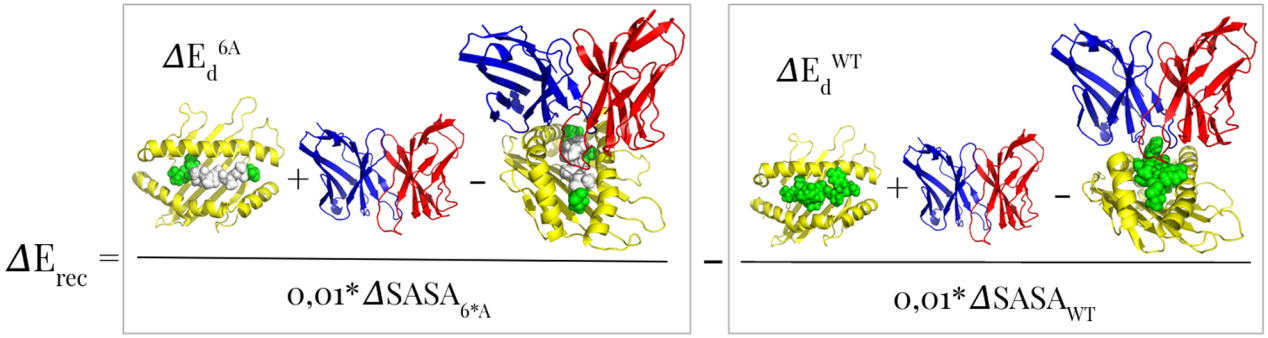

Рисунок 1. Иллюстрация расчёта изменения энергии при распознавании (3, 4) при аланиновых заменах в эпитопе. Красным цветом обозначена β-цепь TCR, синим — α-цепь TCR, жёлтым — MHC, зелёным (сферы) — эпитоп дикого типа (WT), белым (сферы) — аланиновые замены в эпитопе.

со специфическим (5 моделей специфического пептида) и неспецифическими лигандами (5 моделей разных случайных пептидов). Также были рассчитаны разностные значения энергии диссоциации и площади поверхности доступной растворителю с помощью приложения InterfaceAnalyzer. Наибольшая положительная разница в энергии E_{rec} (3) (предлагаемый в рамках данной работы дескриптор — оценка энергии распознавания иммунорецептором пептидного эпитопа) выбиралась для моделей тройных комплексов одного иммунорецептора с одним аффинным и одним случайным антигенным эпитопом (4) (рис. 1).

$$E_{rec,\,i} = \Delta G_i^{pA} - \Delta G_i^{WT},\ [REU/nm^2] \quad (3);$$

$$\Delta E_{rec,\,i} = E_{rec}^{B} - E_{rec}^{NB},\ \text{для } TCR_i,\ [REU/nm^2] \quad (4).$$

Нулевая гипотеза $\mathbf{H_0}$ принимается, если нет разницы, в среднем, между изменением энергии диссоциации при аланиновых заменах в эпитопе (E_{rec}) для структур TCR_i в комплексе с агонистом — высокоаффинным $pMHC^B$ (E_{rec}^B) и не-агонистом — случайным антигеном $pMHC^{NB}$ (E_{rec}^{NB}), то есть $\mu(\Delta E_{rec,i})=0$. Альтернативная гипотеза ($\mathbf{H_1}$) принимается, если модели нативных ТК изменяют свою энергию диссоциации сильнее при аланиновых мутациях в эпитопе, чем случайные ТК, то есть $\mu(\Delta E_{rec,i})>0$.

Разница в изменениях энергии диссоциации при заменах пептидов WT в pA составила $\mu(\Delta E_{rec,i})=7{,}209\times10^{-2}$ REU/nm^2, доверительный интервал составил (0,00018; 0,144) (рис. 2). Значение парного одностороннего t-критерия составило t(200)=4,35 и $p=1{,}1\times10^{-5}$. Следовательно, можно отвергнуть нулевую гипотезу $\mathbf{H_0}$ и принять гипотезу $\mathbf{H_1}$: констатировать наличие положительной разницы в среднем между энергетическими характеристиками лигандных и случайных эпитопов Т-клеточных рецепторов. Значение величины эффекта по Коэну составило 0,41.

Полученные результаты показали, что по сравнению с одним лишь изменением нормированной энергии диссоциации учёт данных вычислительного мутагенеза (WT в pA) позволяет лучше объяснить структурно-энергетическое различие между агонистами и не-агонистами в рамках используемых подходов и моделей.

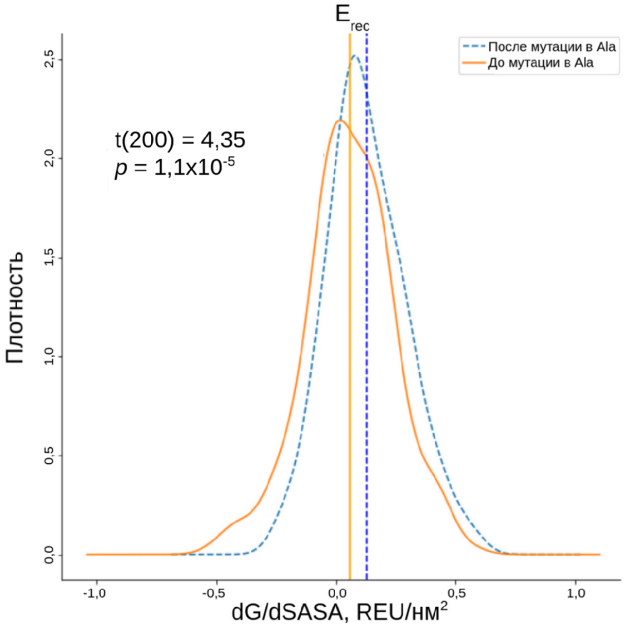

Рисунок 2. Функции плотности вероятности распределений нормированных изменений энергии диссоциации тройных комплексов TCR-pMHC: сплошная линия — до аланиновых замен в эпитопе, прерывистая — после.

Обучение моделей бинарных классификаторов антигенной специфичности иммунорецепторов

Для определения предсказательной способности моделей линейных классификаторов на основе логистической регрессии была создана выборка из 225 TCR × 129 pMHC ТК, из которых были отобраны 12239 ТК (см. раздел “Методы”).

Обучение классификатора на основе геометрической комплементарности

По результатам докинга в программе PatchDock было получено 44249752 линейных преобразования. Для отбора дескрипторов после группировки точек по ТК и усреднения (для некоторых ТК были получены множественные конформации взаимной ориентации TCR-pMHC, для других не получены совсем) был проведён анализ ANOVA (уровень значимости $\alpha=0{,}05$) (табл. 1). На основе отобранных дескрипторов была обучена модель логистической

Таблица 1. Результаты отбора дескрипторов с использованием дисперсионного анализа (ANOVA) для данных, рассчитанных с помощью пакета PatchDock

F	*p*-значение	терм	описание	b*
6,60	1,014E-02	as1	геометрическая комплементарность только на основе антигена	0,0037
59,15	1,57E-14	as2	геометрическая комплементарность только на основе CDR3αβ петель	-0,0043
135,76	3,29E-31	as12	геометрическая комплементарность CDR3αβ и антигена	0,0429
5,17	2,29E-02	score	комплементарность на всём интерфейсе	0,0003

Примечание: F – значение критерия Фишера, b* — коэффициент в модели логистической регрессии для соответствующего дескриптора — компонента оценочной функции PatchDock.

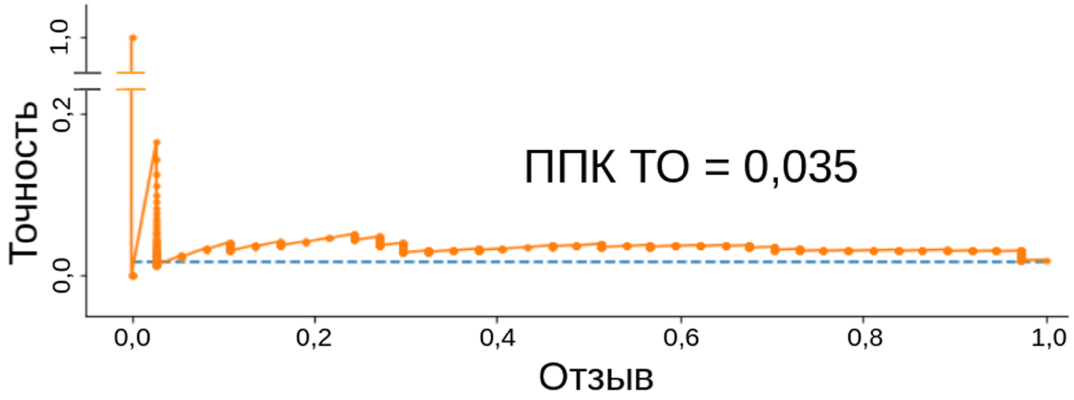

Рисунок 3. Характеристическая кривая точность/отзыв для модели логистической регрессии на основе результатов PatchDock — сплошная линия, для случайного классификатора — прерывистая.

регрессии, при этом набор данных разделяли на обучающий и тестовый в пропорции 3 к 1. Для полученной модели на тестовом наборе были измерены площади под характеристическими кривыми TPR/FPR и точность/отзыв. Значение площадей под кривыми: ROC AUC = 0,75 и ROC PR = 0,035 (рис. 3).

Обучение классификатора на основе полноатомной энергетической функции REF2015

Модель логистической регрессии, описанная выше, была использована для следующего этапа фильтрации ТК с целью минимизации необходимого количества временных затрат на проведение Монте-Карло моделирования выборки полноатомных моделей ТК, используемых для расчёта энергетических дескрипторов с использованием оценочной функции REF2015. Используя полученную характеристическую кривую ROC, было выбрано пороговое значение классификатора равное 0,89, соответствующее значениям TPR=0,9 и FPR=0,5, что привело к отбору 6181 из первоначальных 12239 ТК. Для каждого ТК были взяты 10 моделей с наивысшим соотношением значения геометрической комплементарности к площади образуемого интерфейса. Для всех отобранных моделей был проведён расчёт оптимизации геометрических параметров по Монте-Карло протоколу FastRelax, чтобы избавиться от стерических затруднений, а также по методу FastDesign [16] с заменой ротамеров неякорных остатков пептида на аланил. Для полученных структур были рассчитаны значения полной энергии ТК с помощью оценочной функции REF2015 и на их основе вычислены энергии диссоциации ТК.

С помощью дисперсионного анализа были отобраны дескрипторы (табл. 2): некоторые слагаемые оценочной функции REF2015 [20], рассчитанные для ТК до (WT) и после аланиновых замен в антигене (pA); характеристики заглубляемого при образовании тройного комплекса белок-белкового интерфейса, рассчитанные с помощью приложения InterfaceAnalizer [22]. Таким образом, мы рассматривали следующие энергетически-структурные характеристики: изменение площади поверхности интерфейса, доступной растворителю при диссоциации (5), в том числе отдельно для гидрофобной поверхности интерфейса (6); число остатков, образующих интерфейс; изменение числа свободных акцепторов и доноров водородной связи компонентов ТК между связанным и свободным состоянием комплекса (7); энергию конформации основной цепи на основе карт Рамачадрана (8); часть потенциала Леннарда-Джонса, ответственная за притяжение, сглаженная кубическим полиномом f_{poly} (9), где вес $w_{i,j}$ зависит от количества связей между атомами следующим образом (10); энергетический вклад барьера вращения вокруг пептидной связи, сумма по всем аминокислотным остаткам (a.o.) (11).

Таблица 2. Результаты отбора дескрипторов с использованием дисперсионного анализа (ANOVA) для данных, полученных после оптимизации геометрии моделей ТК по методу FastRelax/FastDesign

F	*p*-значение	терм	описание	b*
7,63	0,006	dSASA_hphobic_WT	Изменение площади гидрофобной части поверхности доступной растворителю при диссоциации ТК (6)	-0,1252
7,63	0,006	omega_pA	Энергетический вклад барьера вращения вокруг пептидной связи, сумма по всем а.о. ТК pA (11)	0,1750
7,16	0,007	omega_WT	Энергетический вклад барьера вращения вокруг пептидной связи, сумма по всем а.о. ТК WT (11)	-0,0816
5,97	0,014	nres_int_WT	Число а.о. на интерфейсе WT	-0,1427
4,68	0,031	dSASA_int_WT	Изменение площади поверхности доступной растворителю при диссоциации ТК (5)	0,0938
4,62	0,032	delta_unsatHbonds_WT	Изменение числа вакантных доноров и акцепторов водородных связей при ассоциации ТК (7)	-0,0775
4,02	0,045	nres_int_pA	Число остатков на интерфейсе pA	0,0880
3,98	0,046	fa_atr_WT	Составляющая потенциала ЛД, ответственная за притяжение (9, 10)	0,0509
3,95	0,047	rama_prepro_pA	Энергия конформации основной цепи, на основе карт Рамачадрана (8)	0,0316

Примечание: F — значение критерия Фишера, b* — коэффициент в модели логистической регрессии REF2015 на основе соответствующего терма данной оценочной функции.

$$\Delta SASA_{int}^{WT} = SASA(TCR) + SASA(pMHC^{WT}) - SASA(TCR - pMHC^{WT}) \quad (5);$$

$$\Delta SASA_{hphobic}^{WT} = SASA_{hphobic}(TCR) + SASA_{hphobic}(pMHC^{WT}) - SASA_{hphobic}(TCR - pMHC^{WT}) \quad (6);$$

$$\Delta N_{unsatHbonds}^{WT} = N_{unsatHbond}^{donors+acceptors}(TCR - pMHC) - N_{unsatHbond}^{donors+acceptors}(TCR) - N_{unsatHbond}^{donors+acceptors}(MHC^{WT}) \quad (7);$$

$$E_{rama\_prepro} = \sum_{r} -ln[P(\varphi_i,\ \psi_i,\ aa)] \quad (8);$$

$$E_{fa\_atr} = \sum_{i,j} w_{i,j}^{conn} \begin{cases} -\varepsilon_{i,j},\ d_{i,j} \le \sigma_{i,j} \\ \varepsilon_{i,j}\left[\left(\frac{\sigma_{i,j}}{d_{i,j}}\right)^{12} - 2\left(\frac{\sigma_{i,j}}{d_{i,j}}\right)^{6}\right],\ \sigma_{i,j} < d_{ij} \ge 0{,}45\ \text{nm} \\ f_{poly}(d_{i,j}),\ 0{,}45\ \text{nm} < d_{i,j} \le 0{,}6\ \text{nm} \\ 0,\ 0{,}6\ \text{nm} < d_{i,j} \end{cases} \quad (9);$$

$$w_{i,j} = \begin{cases} 0,\ n_{i,j}^{bonds} \le 3 \\ 0{,}2,\ n_{i,j}^{bonds} = 4 \\ 1,\ n_{i,j}^{bonds} \ge 5 \end{cases} \quad (10);$$

$$E_{omega} = \sum_{r} ln\left(\frac{1}{6\sqrt{2\pi}}\right) - ln\left(\frac{1}{\sigma_{\omega}(\varphi_r, \psi_r | aa_r)\sqrt{2\pi}}\right) + \frac{[\omega_r - \mu_{\omega}(\varphi_r, \psi_r | aa_r)]^2}{2\sigma_{\omega}^2(\varphi_r, \psi_r | aa_r)} \quad (11).$$

Средние и стандартные отклонения для ω, μ_ω и σ_ω, являются φ,ψ-зависимым и задаются ядерными регрессиями.

Значение ROC AUC для данного классификатора составило 0,65 и ROC PR 0,065 (рис. 4), для сравнения, ROC AUC для классификатора на основе геометрической комплементарности на данной выборке составило 0,6.

Анализ коэффициентов логистической регрессии (табл. 1, 2) позволяет сравнить полученные результаты с предыдущими работами в данной области [9, 20, 21]. Ранее было показано, что среди коэффициентов модели линейной регрессии для зависимости константы диссоциации от слагаемых оценочной функции Rosetta, рассчитанных для структур TCR-pMHC, полученных методом PCA, значимость дескриптора составляющей в потенциале ЛД, отвечающей за притяжение, преобладает над значимостью дескриптора, характеризующего отталкивание [8]. В нашей работе мы также отмечаем данную тенденцию, так как среди дескрипторов, отобранных с помощью дисперсионного анализа, присутствует компонент fa_atr (9), входящий в модель логистической регрессии с положительным коэффициентом. Отрицательное значение коэффициента omega_WT в таблице 2 свидетельствует, что пептидные связи в ТК, образованных с участием агонистов, испытывают меньшее напряжение, чем в комплексах с не-агонистами, что согласуется с опубликованными ранее данными [23]. Положительный коэффициент omega_pA означает, что большее напряжение, возникающее в пептидных связях структуры ТК после замены а.о. центральных позиций антигена на аланин, свидетельствует в пользу того, что ТК образован с участием пептида-агониста. При анализе

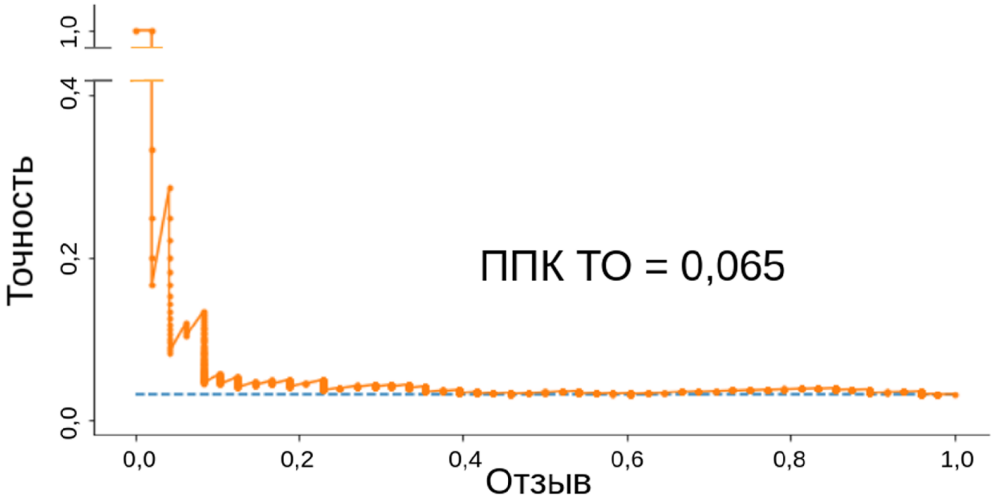

Рисунок 4. Характеристическая кривая точность/отзыв для модели логистической регрессии на основе REF2015 — сплошная линия, для случайного классификатора — прерывистая.

коэффициентов полученной модели логистической регрессии наиболее важным оказывается площадь гидрофобного участка поверхности, заглубляющейся при образовании ТК. Наши результаты позволяют заключить, что при образовании неспецифических взаимодействий, когда геометрическая комплементарность вариабельных участков (CDR3) и антигена меньше (табл. 1), площадь заглублённой поверхности становится больше (табл. 2). Также отмечается статистически значимое отличие в числе заглублённых доноров и акцепторов водородных связей. Хорошо известно, что главной трудностью создания искусственных белок-белковых взаимодействия является создание заглублённых водородных связей [22]. Полученные результаты демонстрируют, что модели ТК с агонистами отличаются лучшим устройством сети водородных связей на белок-белковых интерфейсах.

Описанный нами метод построения TCR-pMHC отличается от ранее опубликованных тем, что он не ограничивается использованием моделирования по гомологии на основании структурного шаблона всего комплекса целиком и позволяет отобрать различные конформации, отличающиеся пространственной ориентацией лиганда относительно рецептора. Предложенный метод достаточно гибко учитывает конформационную подвижность вариабельных участков иммунорецептора и пептида, так как проводится конформационный отбор за счёт явного моделирования возможных конфигураций вариабельных петель до стадии докинга, а также оптимизации геометрических параметров ТК после докинга с учётом внутренних степеней свободы как основной цепи, так и вращательных состояний аминокислотных остатков. В-третьих, предлагаемый метод обладает сравнительно высокой производительностью и легко применим для параллельного запуска на современных компьютерных кластерах.

ЗАКЛЮЧЕНИЕ И ВЫВОДЫ

В ходе проведённого исследования было установлено статистически значимое отличие между энергетическими характеристиками поверхностей взаимодействия, образуемых в процессе распознавания Т-клеточных рецептором пептидов, представленных на поверхности главного комплекса гистосовместимости. На основании полученного набора данных были обучены модели логистической регрессии, позволяющие классифицировать модели тройных комплексов TCR-pMHC на два класса: агонисты и не-агонисты. Стоит отметить, что полученные модели, обладая умеренной предсказательной эффективностью, являются интерпретируемыми и позволяют получить представление о вкладе различных эффектов в распознавание антигенных пептидов иммунорецептором. В перспективе классификаторы, обученные на данных структурного моделирования, могут быть полезны для приоритизации неоантигенов при разработке противоопухолевых вакцин. В особенности, в сочетании с информацией об относительной представленности иммунорецепторов и MHC-связывающей аффинности неоантигенов.

ФИНАНСИРОВАНИЕ

Работа выполнена при финансовой поддержке Российским фондом фундаментальных исследований (проект 19-03-00043) с использованием оборудования Межведомственного суперкомпьютерного центра РАН.

СОБЛЮДЕНИЕ ЭТИЧЕСКИХ СТАНДАРТОВ

Настоящая статья не содержит каких-либо исследований с участием людей или использованием животных в качестве объектов.

КОНФЛИКТ ИНТЕРЕСОВ

Авторы заявляют об отсутствии конфликта интересов.

ЛИТЕРАТУРА

1. *Gopanenko A.V., Kosobokova E.N., Kosorukov V.S.* (2020) Cancers (Basel)., **12**(10), 2879. DOI: 10.3390/cancers12102879.
2. *Reynisson B., Alvarez B., Paul S., Peters B., Nielsen M.* (2021) Nucleic Acids Res., **48**(W1), W449-W454. DOI: 10.1093/NAR/GKAA379.
3. *Baryshnikova M.A., Rudakova A.A., Sokolova Z.A., Burova O.S., Kosobokova E.N., Kosorukov V.S.* (2019) Russ. J. Biother., **18**(4), 76-81.
4. *Bolotin D.A., Poslavsky S., Mitrophanov I., Shugay M., Mamedov I.Z., Putintseva E.V., Chudakov D.M.* (2015) Nat. Methods, **12**(5), 380-381.
5. *Kosorukov V.S., Baryshnikova M.A., Kosobokova E.N., Yakovishina D.Y., Ershova A.S., Pekov Y.A.* (2019) Russ. J. Biother., **18**(3), 23-30.
6. *Lever M., Maini P.K., van der Merwe P.A., Dushek O.* (2014) Nat. Rev. Immunol., **14**(9), 619-629.
7. *Vita R., Overton J.A., Greenbaum J.A., Ponomarenko J., Clark J.D., Cantrell J.R., Wheeler D.K., Gabbard J.L., Hix D., Sette A. et al.* (2015) Nucleic Acids Res., **43**(D1), D405-D412. DOI: 10.1093/nar/gku938.
8. *Borrman T., Cimons J., Cosiano M., Purcaro M., Pierce B.G., Baker B.M., Weng Z.* (2017) Proteins Struct. Funct. Bioinforma., **85**(5), 908-916.
9. *Shugay M., Bagaev D.V., Zvyagin I.V., Vroomans R.M., Crawford J.C., Dolton G., Komech E.A., Sycheva A.L., Koneva A.E., Egorov E.S. et al.* (2018) Nucleic Acids Res., **46**(D1), D419-D427. DOI: 10.1093/nar/gkx760.
10. *Zhang S.Q., Ma K.Y., Schonnesen A.A., Zhang M., He C., Sun E., Williams C.M., Jia W., Jiang N.* (2018) Nat. Biotechnol., **36**(12), 1156-1159.
11. *Zvyagin I.V., Tsvetkov V.O., Chudakov D.M., Shugay M.* (2020) Immunogenetics, **72**(1-2), 77-84.
12. *Springer I., Besser H., Tickotsky-Moskovitz N., Dvorkin S., Louzoun Y.* (2020) Front. Immunol., **11**, 1803. DOI: 10.3389/fimmu.2020.01803.
13. *Ye J., Ma N., Madden T.L., Ostell J.M.* (2013) Nucleic Acids Res., **41**(Web Server issue), 382. DOI: 10.1093/nar/gkt382.
14. *Bender B.J., Cisneros A., Duran A.M., Finn J.A., Fu D., Lokits A.D., Mueller B.K., Sangha A.K., Sauer M.F., Sevy A.M. et al.* (2016) Biochemistry, **55**(34), 4748-4763.
15. *Gowthaman R., Pierce B.G.* (2018) Nucleic Acids Res., **46**(W1), W396-W401. DOI: 10.1093/nar/gky432.
16. *Khatib F., Cooper S., Tyka M.D., Xu K., Makedon I., Popović Z., Baker D., Players F.* (2011) Proc. Natl. Acad. Sci. USA, **108**(47), 18949-18953.
17. *Leaver-Fay A., Tyka M., Lewis S.M., Lange O.F., Thompson J., Jacak R., Kaufman K., Renfrew P.D., Smith C.A., Sheffler W. et al.* (2011) Methods Enzymol., **487**(C), 545-574.
18. *Stein A., Kortemme T.* (2013) PLoS One, **8**(5), e63090. DOI: 10.1371/journal.pone.0063090.
19. *Nerli S., Sgourakis N.G.* (2020) Front. Med. Technol., **2**, 553478. DOI: 10.3389/fmedt.2020.553478.
20. *Alford R.F., Leaver-Fay A., Jeliazkov J.R., O'Meara M.J., di Maio F.P., Park H., Shapovalov M.V., Renfrew P.D., Mulligan V.K., Kappel K. et al.* (2017) J. Chem. Theory Comput., **13**(6), 3031-3048.
21. *Schneidman-Duhovny D., Inbar Y., Nussinov R., Wolfson H.J.* (2005) Nucleic Acids Res., **33**(SUPPL. 2), W363. DOI: 10.1093/nar/gki481.
22. *Stranges P.B., Kuhlman B.* (2013) Protein Sci., **22**(1), 74-82.
23. *Borrman T., Pierce B.G., Vreven T., Baker B.M., Weng Z.* (2020) Bioinformatics, **36**(22-23), 5377-5385. DOI: 10.1093/bioinformatics/btaa1050.

Поступила в редакцию: 15. 04. 2021.
После доработки: 08. 05. 2021.
Принята к печати: 09. 05. 2021.

IN SILICO SPECIFICITY DETERMINATION OF NEOANTIGEN-REACTIVE T-LYMPHOCYTES

***A.E. Kniga*[1,2]*, *I.V. Polyakov*[1,2], *A.V. Nemukhin*[1,2]**

[1]M.V. Lomonosov Moscow State University,
GSP-1, 1 Leninskie Gory, Moscow, 119991 Russia; *e-mail: kniga.ae@gmail.com
[2]N.M. Emanuel Institute of Biochemical Physics RAS,
4 Kosygina str., Moscow, 119334 Russia

Effective personalized immunotherapies of the future will need to capture not only the peculiarities of the patient's tumor but also of his immune response to it. In this study, using results of *in vitro* high-throughput specificity assays, and combining comparative models of pMHCs and TCRs using molecular docking, we have constructed all-atom models for the putative complexes of all their possible pairwise TCR-pMHC combinations. For the models obtained we have calculated a dataset of physics-based scores and have trained binary classifiers that perform better compared to their solely sequence-based counterparts. These structure-based classifiers pinpoint the most prominent energetic terms and structural features characterizing the type of protein-protein interactions that underlies the immune recognition of tumors by T cells.

Key words: neoantigen; molecular docking; protein-protein interactions; comparative modeling; machine learning; TCR-pMHC

Funding. This work was financially supported by the Russian Foundation for Basic Research (project no. 19-03-00043) and conducted using the equipment of the Joint Supercomputer Center of the Russian Academy of Sciences.

Received: 15.04.2021, revised: 08.05.2021, accepted: 09.05.2021.