Математическая биология и биоинформатика
2023. Т. 18. № 2. С. 621–645. doi: 10.17537/2023.18.621

========== **БИОИНФОРМАТИКА** ==========

Статистическая модель предсказания сайтов связывания TALEN с ДНК на основе скользящего среднего

Тетуев Р.К., Назипова Н.Н.

Институт математических проблем биологии РАН – филиал ИПМ им. М.В. Келдыша РАН, Пущино, Россия

Аннотация. В работе представлен новый подход *in silico* предсказания всех вероятных сайтов связывания искусственных ферментов рестрикции TALEN, основанный на модели экспоненциально взвешенного скользящего среднего. Подход реализован в виде онлайн-сервиса TANDIS с разделением вычислительного процесса на четыре равномощных потока за счёт естественного параллелизма на уровне данных. Прямая верификация качества предсказания проводилась путем сравнения с результатами *in vitro* экспериментов, а косвенная верификация – сравнением с результатами предсказаний аналогов: TALE-NT, Galaxy/TALENoffer. Анализ результатов показал, что по качеству предсказания TANDIS и TALENoffer сравнимы между собой, но простота нашей математической модели позволяет значительно ускорить вычислительный процесс, в то время как качество результатов TALE-NT значительно уступает обоим этим подходам. Алгоритм TANDIS основан на прямом моделировании физического взаимодействия белка и двойной спирали ДНК, этот подход может оказаться полезным для понимания процессов, протекающих в клетке на микробиологическом уровне и иметь важное значение для дальнейшего развития генной инженерии.

Ключевые слова: *редактирование геномов, TALEN, экспоненциально взвешенное скользящее среднее, in-silico предсказание сайтов связывания, программное обеспечение.*

ВВЕДЕНИЕ

Редактирование генома является активно развиваемой областью науки и практики, в которой особенно важна роль «геномных ножниц», используемых для специфичных модификаций ДНК любого организма путём разрезания ДНК нитей в некоторой особой целевой позиции, необходимой исследователям или практикам.

Еще в 1990-е годы мегануклеазы и нуклеазы цинковых пальцев (Zinc Finger Nucleases, ZFNs) заложили основу концепции редактирования генома. В 2002 году это развитие завершилось прорывным достижением – созданием первого в мире организма с отредактированным геномом [1]. В 2010 году были разработаны новые мегануклеазы TALEN (Transcription Activator-Like Effector Nuclease), представляющие собой слияние эффектора, подобного активатору транскрипции (TALE), и каталитического домена эндонуклеазы рестрикции FokI. Это были первые конструкции, которые можно было спроектировать и изготовить для воздействия на любой конкретный локус генома с высокой точностью и высокой эффективностью. TALEN стали применять для редактирования геномов сельскохозяйственных культур, домашнего скота, а также производства модельных и немодельных организмов. Технология TALEN стала первым практическим инструментом редактирования генома, который спас человеческую жизнь, вылечив рак в 2015 году [2], и вывел на рынок первую культуру с отредактированным геномом в 2019 году [3]. Примечательно, что, хотя ZFN и мегануклеазы сложнее

использовать, они по-прежнему используются биотехнологическими компаниями, обладающими соответствующими навыками.

С 2012 года получили широкое распространение многообещающие CRISPR/Cas системы редактирования генома. Основным преимуществом CRISPR/Cas перед технологией TALEN было существенная простота использования, что позволило лабораториям по всему миру принять редактирование генома в качестве дешевой рутинной технологии. Было разработано множество вариантов CRISPR-систем, имеющих различные приложения [4, 5]. Однако ограничения, связанные с высокой частотой нецелевой активности, обнаруженные у CRISPR-систем, не позволяют забыть о высокоспецифичных мегануклеазах TALEN и ZFN, особенно когда речь идет о биомедицинских приложениях.

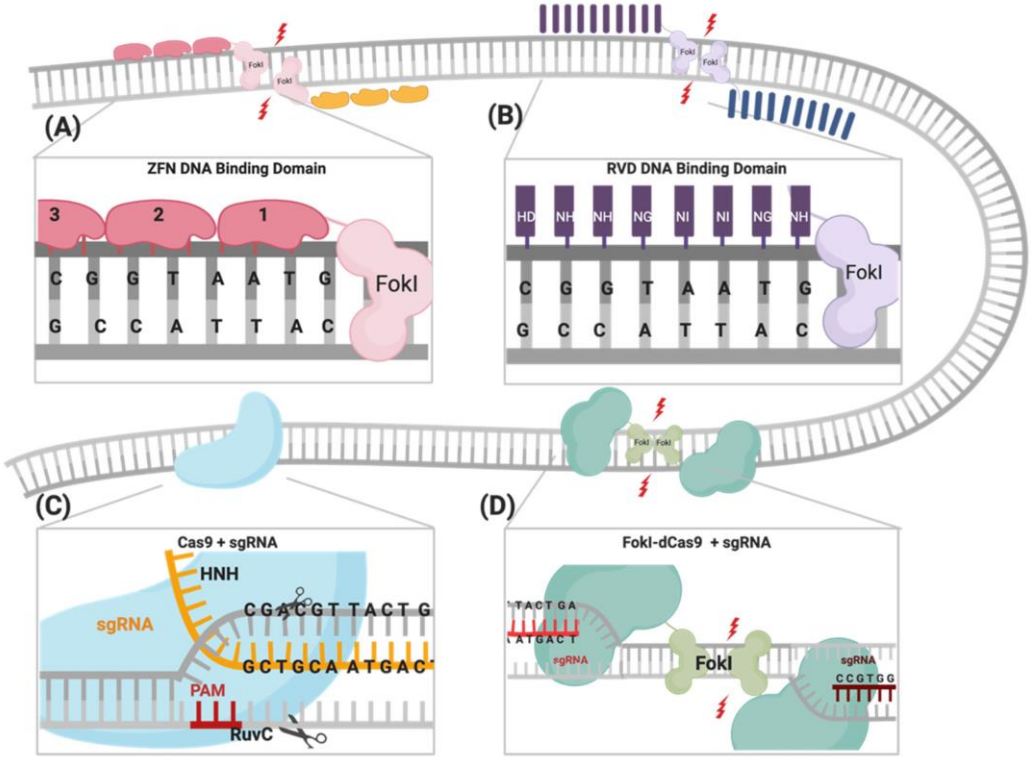

Рис. 1. На иллюстрации, заимствованной из [6], показаны для сравнения принципы работы базовых технологий редактирования ДНК. (**A**) Нуклеазы с цинковыми пальцами (ZFN) связываются с целевой последовательностью посредством 3-4 белков с цинковыми пальцами, и (**B**) TALEN связывается посредством повторяющихся вариабельных остатков (RVD). В обеих этих платформах индуцируются двухцепочечные разрывы ДНК за счет гетеродимеризации и каталитической активности FokI. (**C**) CRISPR-Cas9 связывается с ДНК-мишенью, комплементарной направляющей РНК (sgRNA), положение мишени определяется соседством с мотивом PAM, а каталитическая активность осуществляется доменами HNH и RuvC нуклеазы Cas9. (**D**) FokI-dCas9 связывается с сайтом-мишенью посредством двух sgRNA, которые расположены в ориентации PAM-out, а каталитическая активность опосредована димеризацией эндонуклеазы FokI.

Делаются попытки разработки гибридных стратегий высокоспецифичного редактирования генома, например, CRISPR-FokI-dCas9 (fdCas9). РНК-ориентированная нуклеаза FokI (RFN) была впервые представлена и проверена в 2014 году. dCas9 – это т. нзв. неактивная «мертвая» форма Cas9. Система fdCas9 получается путем связывания каталитического домена эндонуклеазы FokI с неактивной нуклеазой Cas9 и требует пары направляющих РНК, которые связываются со смысловой и антисмысловой цепями ДНК в ориентации PAM-out. При дизайне двух сайтов нацеливания sgRNA, фланкирующих целевой ген, нужно выбрать правильную длину спейсера между двумя мишенями, две последовательности PAM, которые должны находиться на противоположных цепях, обращенными наружу от концов спейсера. Димеризация доменов FokI генерирует

двухцепочечные разрывы ДНК, которые активируют механизм репарации ДНК и приводят к редактированию генома. Утверждается [6], что все сконструированные варианты fdCas9 продемонстрировали многообещающую активность по редактированию генов в клетках человека по сравнению с другими платформами. На рисунке 1 схематически показаны принципы работы перечисленных технологий редактирования генов.

Структура и функции TALE и TALEN

Эффекторы, подобные активаторам транскрипции (TALE), представляют собой белки фитопатогенных бактерий *Xanthomonas*. Их вводят в растительные клетки, где они локализуются в ядре, связываются с целевыми промоторами и индуцируют экспрессию генов. Белки TALE содержат три функциональных домена. Их N-концевой домен несет в себе сигнал бактериальной секреции и неспецифическую ДНК-связывающую активность, необходимую для общего сродства белка к ДНК (рис. 2) [7]. Их С-концевой домен содержит интерфейс взаимодействия растительного транскрипционного фактора IIA, два функциональных сигнала ядерной локализации и домен кислотной активации [8]. Ключом к их специфическому и программируемому связыванию с ДНК является центральная область повторов, которая состоит из переменного количества тандемных повторов, обычно состоящих из 33–35 аминокислот [9]. Две аминокислоты, которые определяют специфичность ДНК TALE, расположены в положениях 12 и 13 каждого повтора, в повторяющихся вариабельных аминокислотных остатках (Repeat Variable Diresidue, RVD) (рис. 2). Были расшифрованы особенности всех возможных комбинаций RVD, выявлены очень специфические, распознающие только один нуклеотид, а также более гибкие, допускающие два, три или все четыре нуклеотида [10, 11]. Перестраивая повторы, специфичность связывания ДНК TALE может быть изменена по желанию.

Кристаллическая структура TALE показала, что они образуют правостороннюю суперспиральную структуру, охватывающую ДНК. Каждый повтор образует две альфа-спирали, которые соединены петлевой областью, в которой расположены RVD [12, 13]. Кроме того, TALE содержат четыре вырожденных, или неканонических, повтора (называемых –3, –2, –1 и 0) в своей N-концевой области [7], причем повтор –1 распознает дополнительный тимин, который предшествует участку, связанному с повторами TALE. Подход направленной эволюции позволил получить варианты N-концевого домена TALE, которые распознают все основания, что упрощает позиционирование TALE и TALEN. Специфичность повтора может быть выбрана свободно, чтобы соответствовать любой желаемой целевой последовательности, однако рекомендуется включать по крайней мере 2 или 3 C или G в целевой сайт TALE или TALEN и включать RVD HD и NN соответственно. Оба считаются сильными RVD, и TALE без таких сильных RVD может быть крайне неэффективным [14]. Поскольку количество повторов в TALEN можно регулировать относительно свободно, можно использовать пару TALEN для точного разрезания нужного нуклеотида в целевой последовательности ДНК.

В обзоре [15] содержится подробное описание возможностей TALE, предлагаются две уникальные особенности, которые расширяют диапазон применения этой технологии. Первая основана на использовании редко встречающихся в природе повторяющихся единиц аберрантной длины. Если один из этих аберрантных повторов включен в массив повторов, TALE или TALEN могут распознавать не только нормальные целевые последовательности, но и последовательности с делецией в 1 пару оснований вблизи аберрантного повтора. Вторая особенность основана на способности некоторых PVD различать нуклеотиды с разными состояниями метилирования. Например, естественный RVD HD распознает цитозин, но не способен распознавать 5-метилцитозин (5mC) или 5-гидроксиметилцитозин (5hmC), тогда как RVD N* распознает метилированные и неметилированные цитозины. TALE даже способны различать две целевые последовательности, которые отличаются только одним метилированным нуклеотидом, и, таким образом, TALE и TALEN можно использовать для манипуляций с генами,

зависимых от метилирования.

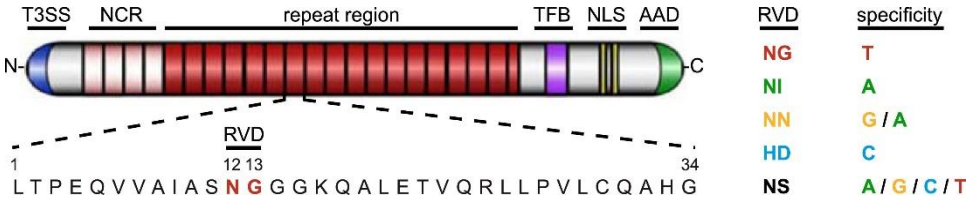

Рис 2. Архитектура TALE. N-концевая область содержит сигнал секреции типа III (T3SS) и четыре неканонических повтора (NCR), С-концевая часть – сайт связывания транскрипционного фактора (TFB), два сигнала ядерной локализации (NLS) и домен кислотной активации (AAD). Каждая TALE также содержит область повторов с различным количеством высококонсервативных повторов из 33–35 аминокислот, расположенных тандемно. Показана аминокислотная последовательность консенсусного повтора TALE с 34 аминокислотами и выделены аминокислоты, ответственные за специфичность TALE (RVD). Также показаны пять наиболее часто используемых RVD и нуклеотиды, с которыми они связываются. Заимствовано из [15]

С помощью экспериментов с одной молекулой было обнаружено, что TALE используют неожиданный механизм для поиска сайтов-мишеней ДНК: они обволакивают ДНК и выполняют быстрый, одномерный, невращательный и беспристрастный поиск. Это резко контрастирует с другими ДНК-связывающими белками, специфичными для последовательности. Обычно они используют либо «прыжковый» механизм, который характеризуется частыми де- и реассоциациями белка с ДНК, либо «скользящий» механизм, при котором белок вращается вокруг ДНК, следуя по ее основной бороздке. Скорость TALE превышает скорость типичных «скользящих» ДНК-связывающих белков, что позволяет им быстро идентифицировать совпадающие целевые последовательности даже в больших геномах [15].

TALE используют рыхлую конформацию в состоянии без ДНК и в процессе поиска при движении по ДНК, но переключаются на более конденсированную конформацию при входе в режим распознавания и связывания. Остается неясным, может ли этот переход между расслабленным и конденсированным режимами начаться из любой точки внутри области повторения или это направленный процесс. N-конец TALE с его неспецифическим сродством к ДНК, его важность для первоначального связывания TALE с ДНК предполагает, что первые повторы являются решающим фактором для первоначального распознавания целевой последовательности и могут инициировать конденсацию [15].

TALEN – это искусственный белок, состоящий из трёх доменов: системы секреции III типа (T3SS), ряда TALE-повторов и эндонуклеазы рестрикции Fok1. Причём, если первые два домена, T3SS и TALE-повторы, присутствовали в нативном белке AvrBs3 из *Xanthomonas citri* патовар *citri*, где и были впервые обнаружены TALE, то рестриктаза Fok1 взята из белка другой бактерии *Flavobacterium okeanokoites* взамен нативного домена активации [16]. Химерный белок TALEN способен использовать T3SS для проникновения в клетку и TALE-повторы для поиска в геноме специфичного сайта связывания, но не вызывает ложную и чрезмерную транскрипцию определённых генов, вызывая гибель клеток хозяина (как это делает *Xanthomonas citri*), а лишь проделывает надрез одной нити двухцепочечной ДНК, что определяется перепрограммированием TALE-повторов. Следует понимать, во-первых, место надреза будет где-то на участке длиной в 10–30 пар нуклеотидных оснований (н.п.) от сайта связывания в сторону 3'-конца, а во-вторых, это будет именно надрез, когда разрывается только одна нить двойной спирали ДНК. В некоторых экспериментах бывает нужно осуществить одновременный разрез обеих нитей ДНК. Тогда используют вторую TALEN, которая должна найти на комплементарной нити ДНК определённый сайт связывания, выбранный так, чтобы надрез одной нити оказался поблизости от надреза другой, или, в идеале, ровно под ним, что почти гарантирует осуществление полного разрыва в этом месте.

Каждая из двух нуклеаз в этом случае должна обвить двойную спираль ДНК и скользить вдоль большой бороздки до специфического сайта связывания, обрамляя целевой участок слева (5'-конец) и справа (3'-конец). При этом специфичность связывания с целевым сайтом задаётся набором из особых участков центрального домена белка TALEN (RVD). Таким образом, связывание группы последовательно расположенных нуклеотидов ДНК и соответствующей группы RVD работает по принципу "ключ-замок" и вероятность этой связи зависит от правильного выбора RVD для соответствующих позиций.

Актуальность создания упрощенной модели предсказания сайтов связывания

К недостаткам всех существующих технологий редактирования геномов относится способность агентов связывания к неспецифичному взаимодействию с геномом (*off-target activity*). Не лишена этого недостатка и технология TALEN, цепочка RVD способна к эффективному связыванию не только с целевым сайтом, но и с другими (нецелевыми) участками генома, с которыми она связывается с ошибками. Хотя случаи неспецифического связывания происходят реже специфических взаимодействий TALEN с ДНК, но, например, в случаях медицинских применений технологии это нежелательно вовсе. Полностью исключить нецелевую активность невозможно, однако можно минимизировать риски, поскольку почти всегда существует несколько вариантов выбора участков для целевого TALEN-ДНК связывания. Задача выбрать тот вариант, который имеет наименьшее число побочных эффектов.

Проверка и исключение потенциально опасных TALEN требуют доклинических исследований и изначально они сводились к перебору всех вариантов в ходе довольно дорогостоящих *in vitro* экспериментов методом SELEX [17, 18], а затем отбрасывались неудачные варианты. Теперь появилась возможность снижать расходы на разработку TALEN и многократно сокращать число вариантов, т. к. на основании *in vitro* результатов можно создавать математические модели для описания поведения TALEN [19].

Исторически первые попытки поиска потенциальных неспецифических сайтов связывания с ДНК разработаны на основе системы оценок индивидуальных пар RVD:нуклеотид и заданы весовыми таблицами соответствующих подстановок. И несмотря на первые успехи этого подхода, на практике возникали ситуации, когда некоторые участки с меньшей степенью схожести проявляли более частое связывание с TALEN по сравнению с *in silico* рассчитанной оценкой вероятности связывания. Последующие уточнения матрицы подстановок никак не улучшали ситуацию, пока в работе [20] не выдвинули предложение о решающем влиянии взаимного расположения ошибок. Эта идея в [20] описана в виде марковской цепи в самом общем виде, когда степень индивидуального связывания RVD:нуклеотид зависит от всего контекста (от соседних нуклеотидов и RVD), а также от окружения потенциального сайта связывания. В модель оценок вошли более полусотни различных параметров, значения которых установлены в ходе модельных экспериментов. Такая усложненная модель оправдала себя на практике, предсказав множество нежелательных сайтов связывания из числа пропускаемых более ранним и простым ПО. Однако модель [20] не даёт разумной биологической интерпретации полученным успешным результатам и, несмотря на ее широкое применение, становится понятно, что марковская цепь в общем виде не приближает нас к пониманию механизма связывания.

Мы предлагаем упрощённую наглядную гипотезу, согласно которой процесс связывания TALEN и ДНК похож на работу застежки-молнии, а каждой ошибке связывания соответствует сломанный зубчик. Ясно, что десять удалённых зубчиков, расположенных порознь, всего лишь понизят надёжность молнии, однако 3-4 сломанных зубчика, сгруппированных вместе, приведут к полной потере функционала замка-молнии.

Этот принцип легко формализовать с помощью уравнения свёртки, например, описав оценки на основе экспоненциального скользящего среднего, применяя их к имеющимся

табличным оценкам. В результате развития этой идеи авторам удалось развить принципиально новый метод предсказания сайтов связывания TALEN и разработать программное обеспечение TANDIS. Сравнение с экспериментальными данными показало, что такая аналогия более чем уместна, так как она позволяет сократить число параметров модели до пяти, упростить и значительно ускорить процесс предсказания, не уступая при этом по качеству аналогам с более сложными моделями. С точки зрения физики подобная модель также может оказаться полезной для лучшего понимания этого микробиологического процесса.

МАТЕРИАЛЫ И МЕТОДЫ

Белковый домен из TALE повторов способен обвить двойную цепь ДНК, накрывая небольшой её участок из такого же количества нуклеотидных оснований, по одному нуклеотиду на повтор. ДНК представляется обычно в виде текстовой строки, описание каждого TALE-повтора сокращается до двух представителей от каждого повтора (RVD), определяющих эффективность установления связи между конкретным повтором и соответствующим нуклеотидным основанием на цепи ДНК. Анализ природных пар из аминокислотных оснований RVD позволил выделить четыре основных типа: NI, HD, NN, NG, а на основе экспериментальных наблюдений установлено правило приоритетного соединения путём образования водородных связей между парой аминокислот и одним из нуклеотидов: {NI:A, HD:C, NN:G, NG:T}.

Предсказание наиболее вероятных мест связывания TALE с ДНК во многом напоминает текстовый поиск по шаблону. Действительно, если, например, природный белковый домен из пяти TALE повторов имеет описание HD-HD-NI-NN-NG, то он явно запрограммирован на связывание с сайтом-мишенью CCAGT. Но нормальное функционирование природных механизмов имеет множество допущений, что придаёт им пластичность и устойчивость к изменениям условий среды. Это выражается в допустимости альтернативных связей – каждая пара RVD иногда может связаться не со своим приоритетным нуклеотидным основанием. Так, пара NN обычно образует связи с гуанином (NN:G), но на каждый третий случай приходится соединение с аденином (NN:A). Альтернативные, более слабые связывания существуют для каждой пары RVD. Статистический анализ TALE белков показал, что в природе встречаются вообще любые сочетания RVD и нуклеотидов, просто некоторые возникают очень редко.

Вполне естественно представлять количественную оценку возможности связи как оценку вероятности связи, что сильно упрощает подход. Все детали процесса связывания TALE с ДНК на молекулярном уровне неизвестны, но недавно выяснилось, что белок опоясывает двойную цепь, ложась в большую бороздку, и скользит в сторону 5'-конца, пока участок под белком на верхней цепи ДНК не окажется подходящего состава – тогда происходит закрепление и активируется следующий домен белка. У природных TALE это активатор транскрипции, а в искусственных (химерных) белках – обычно эндонуклеаза рестрикции. При этом белок касается цепи ДНК именно своими RVD парами, как бы «прощупывая» её содержимое (такое поведение предсказано, вычислительно смоделировано, доказано экспериментально), но всё равно неясно, каков вклад каждой из водородных связей, и по какому правилу из них складывается успешность при «неидеальном» совпадении пар.

Мы начали свое исследование с анализа статистики для точечных связей, чтобы потом предложить модель для интегральной оценки и естественным образом прийти к оценке вероятности связи. Воспользуемся логарифмом отношения шансов, т.к. этот статистический показатель давно успешно применяется не только в медицинских, социологических, но и в биоинформатических исследованиях, например, при получении весовых матриц замен, используемых для оценки сходства аминокислотных последовательностей (PAM, BLOSUM).

Мы разработали процедуру анализа природных TALE, используя актуальные данные Genbank. Технически подобный анализ состоит из пяти шагов:

1) Скачать из Genbank все известные геномы рода *Xanthomonas*, используя запрос https://www.ncbi.nlm.nih.gov/genome/?term=Xanthomonas.

2) Локализовать в геномах участки тандемных повторов, у которых мотивы имеют длину в диапазоне 99–102 н. п.

3) Осуществить трансляцию полученных участков в аминокислотные последовательности, причём во всех шести рамках считывания, используя не стандартный генетический код, а специализированный генетический код для бактерий, архей и прокариотических вирусов.

4) Выбрать из шести рамок ту, что содержит повторы с фрагментами GGKQALET и выровнять их, считая началом мотива участок LTP.

5) Подсчитать в каждой позиции распределение частот по всем аминокислотам, а затем вывести результаты в виде ярусной диаграммы (гистограммы с накоплением), как на рисунке 3.

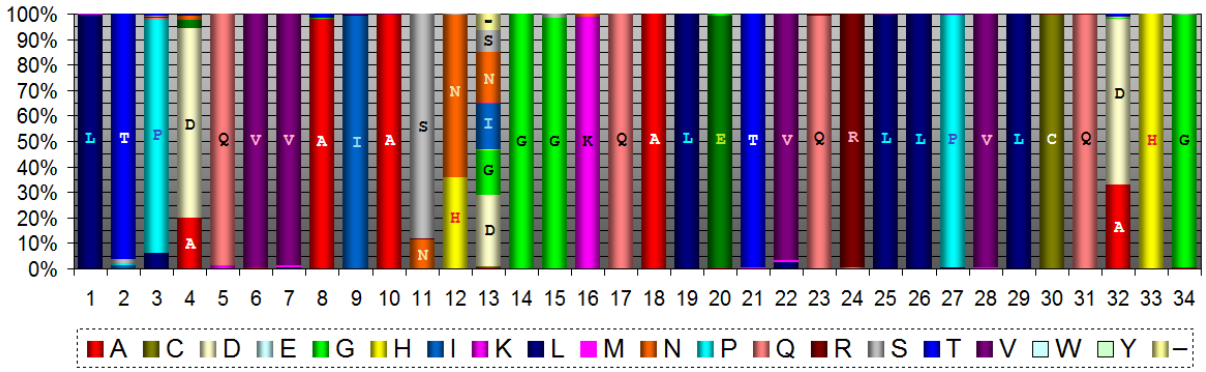

Рис. 3. Пример ярусной гистограммы для TALE-повторов. Она наглядно демонстрирует высокую вариабельность в позициях 12–13 и консервативность длины повторов (33–34), что было замечено в ранних работах, например, в [9].

На построенной нами ярусной диаграмме (рис. 3) видно, что множество RVD состоит из всех возможных парных сочетаний аминокислот, обнаруженных в позициях 12–13, и отнюдь не ограничивается указанными выше четырьмя парами. Легко заметить вариативность и на 11-ой позиции (перед RVD), причины возникновения которой пока неясны. Известно, что во всех природных TALE белках аспарагин (N) замещает в этой позиции серин (S) тогда, когда в этом же TALE повторе за ним далее следуют NN в позиции RVD. Но не наоборот, как показал статистический анализ, перед NN с равной вероятностью может оказаться и серин, и аспарагин. Возможно, что выбор между S и N как-то влияет в природе на селективность данного RVD – в любом случае рекомендуется для связи RVD:G применять не NN, а NH, несмотря на то, что это гораздо более редкий в природе вариант RVD [21].

Выявлены зависимости между RVD и выбором аминокислот на 4-ой и 32-ой позициях соседних TALE повторов, на основании чего делается предположение о регуляции длины повторов за счёт таких вариаций в местах, где это возможно сделать без потери функциональности [21]. Вспомним, что в природе белок наматывается на ДНК, а каждая из аминокислот имеет свой физический размер, т. е. одни RVD шире других, что в сумме приводит к их существенным смещениям относительно целевых нуклеотидов – похоже, выбор между короткой аминокислотой аланином (A) и крупной аспарагиновой кислотой (D) служит белку компенсаторным механизмом, позволяющему ему улечься в жёсткие рамки большой бороздки ДНК [22].

Несмотря на то, что правильный подбор аминокислот в искусственных TALE белках в 4-ой, 11-ой и 32-ой позициях повысил бы, очевидно, эффективность его связывания с

ДНК, на практике спроектировать искусственный белок «совсем как в природе» сложно – исследователи не располагают сотней лет для проведения селекции и средствами для миллионы натурных экспериментов. Конечно же, теоретически можно воспользоваться методами молекулярной динамики, спрогнозировать поведение каждой из версий TALE белка, а затем отобрать наилучшую из них, но и это достаточно дорого. На практике же пока этот вопрос решается довольно просто. В рамках ряда коммерческих закрытых работ проверена эффективность искусственных TALE в случае, когда не-RVD вариации заполнены согласно повторяющимся паттернам. Например, для 32-ой позиции среди рассматриваемых паттернов наилучшим (в среднем) оказалось простое чередование аланина и аспарагиновой кислоты вида «DAADDAAD…» [23]. На основании проведённой работы появляются новые коммерческие решения и наборы инструментов для получения эффективных искусственных TALEN вида Platinum Gate TALEN Kit [24].

Вернёмся к списку исследуемых бактерий *Xanthomonas citri*, *Xanthomonas cannabis*, *Xanthomonas oryzae* и других – в названиях вида этих бактерий упоминаются растения, которые инфекции поражают, вызывая характерную по виду чёрную гниль на листьях и плодах. Разрастание такой бактериальной пятнистости вызывает отмирание тканей, а после попадания в грунт патогены продолжают свой жизненный цикл, переносясь далее на следующие растения благодаря циркуляции воды. Но важно вспомнить, бактерии не тратят собственных сил на разрушение живой клетки, а заставляют её саму это сделать, вызывая принудительную и аномально интенсивную экспрессию некоторого целевого гена самого растения, что и приводит к гибели клетки. Однако у каждой инфекции свой целевой ген и свой участок перед геном, который узнаётся распознающим доменом из TALE повторов – по повтору на один нуклеотид. Здесь несложно заметить, что *Xanthomonas*, которые атакуют дикие растения, редко насчитывают более 15-20 повторов TALE, однако патогены культурных видов вынуждены расширять их число до 30 и более повторов, что безусловно является очевидным проявлением *эволюционной гонки вооружений* (протяжённые домены лучше находят нужные участки в ДНК растения и целевой ген за ним даже при наличии ряда мутаций, которые растения, в свою очередь, культивируют, пытаясь избавиться от этой инфекции) [25].

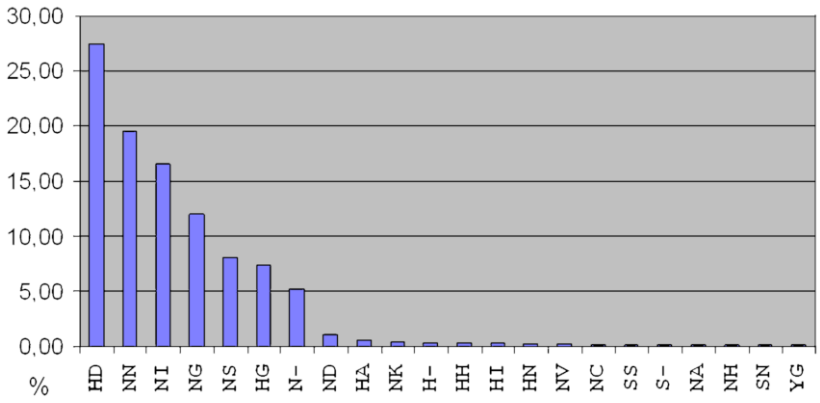

Рис. 4. Частота встречаемости RVD пар в природе, полученная на основании статистического анализа естественных TALE белков.

Тогда очевидно, что, выделив TALE повторы из ДНК, например, *Xanthomonas oryzae* и имея секвенированный геном риса, *Oryza sativa*, не составит труда найти в ДНК растения тот самый ген и целевой участок перед ним, в который и метит TALE данного патогена. При этом станет видно, каким RVD соответствуют в природе какие нуклеотиды, а значит переставляя RVD местами возможно «перепрограммировать» белки на связь с другим порядком нуклеотидов, т. е. с любым заданным участком ДНК, причём это сработает не только в растениях, но и в живых клетках человека. Но так как соответствие

RVD:нуклеотид, строго говоря, не является взаимооднозначным, или даже синонимичным, то имеет смысл распространить исследование на все такие патогены, собрав статистику по частоте встречаемости всех попарных сочетаний RVD, найденных в природе (рис. 4), и четырёх основных нуклеотидов (A, C, G, T).

Следует отметить, что проделать подобное исследование по всем патогенам из рода *Xanthomonas* является грандиозной задачей, выполнение которой нелегко осуществить даже силами двух-трёх научных институтов, однако большим подспорьем в этом деле оказался тот факт, что в каждой стране большое сельскохозяйственное значение имеют разные культуры и очень скоро после открытия принципа «кодирования» TALE белков [16] начало появляться множество работ, посвящённых обнаружению целевых участков для практически всех значимых для человека злаков, фруктов и овощей. Основные выдержки из этих работ любезно предоставлены авторами статьи [26]. В нашей работе используются только самые надёжные сайты связывания.

Теперь, имея в наличии необходимые данные, остаётся только провести их анализ и поместить все обнаруженные случаи соответствия RVD:нуклеотид в таблицу 1 (совсем редкие RVD остались проигнорированы).

Таблица 1. Частота связывания диаминокислот RVD с нуклеотидами в природе

Нуклеотиды	NI	HD	NN	NG	NS	N-	HG	H-	IG
A	118	24	14	7	45	3	0	0	0
C	22	150	12	17	18	30	3	0	0
G	3	0	33	3	5	3	0	0	0
T	4	7	3	108	4	9	13	2	2
Total	147	181	62	135	72	45	16	2	2

На основании собранных данных следует предложить количественную оценку связи RVD:нуклеотид с точки зрения её предпочтительности и, как замечено выше, для этого наиболее естественным будет использовать логарифм отношения шансов.

К примеру, если предположить, что некая RVD в природе встретилась 40 раз, будет ли это говорить о высокой значимости такой RVD? На самом деле важно отношение шансов между ожиданием сочетания RVD:A или RVD:C, или RVD:G, или RVD:T. Поэтому вычисляется среднее значение от числа встречаемости по всем нуклеотидам для конкретного RVD m=40/4=10. Если частота сочетаний с этим RVD одинакова для всех нуклеотидов, то получим одинаковую оценку для всех: $\ln(10/m) = 0$, что означает абсолютную нечувствительность RVD к выбору нуклеотида – он не вносит никакого вклада в общую связь белка с ДНК, ни положительного, ни отрицательного. Но если есть предпочтения, если частоты связей распределены неравномерно, например, так:

Частоты (RVD, X) = {A: 5, C: 5, G: 20, T:10},

тогда для связей RVD:A и RVD:C получим оценку: $\ln(5/10) \approx -0.69$, а для RVD:G будет $\ln(20/10) \approx +0.69$, в то время как для RVD:T она равна $\ln(10/10)=0$:

Баллы (RVD, X) = {A: – 0.69; C: – 0.69; G: +0.69; T: 0}.

Так мы получили оценки, хорошо согласующиеся с интуитивным пониманием приоритетности связей: самая предпочтительная связь с гуанином (положительная оценка), а не с аланином и цитозином (отрицательные оценки), а вот связь с тимином нежелательна, но допустима (промежуточная оценка). Баллы – это оценки соответствия отдельных пар RVD:нуклеотид, наиболее вероятными для связи участками можно будет считать те, которые наберут наибольшую сумму баллов. При подобном моделировании система намеренно упрощается, и игнорируется масса факторов, включая состав, количество, даже взаимное расположение индивидуальных связей, самые первые инструменты для *in silico* предсказания участков связывания построены по принципу

суммирования баллов [27].

Разработанная нами методика позволяет разрабатывать искусственные TALE эффекторные белки с заданными свойствами, причём ещё на стадии выбора участков для связи система сканирует весь геном целевого организма и предостерегает разработчиков о возможной активности выбранной конфигурации в нецелевых локусах. Результаты сканирования геномной ДНК представляет собой список из участков-кандидатов, отсортированный по убыванию общей оценки, в первой строке – целевой участок с наибольшей суммой баллов. Опасаться следует ситуации, когда оценки в следующих строках окажутся достаточно высокими, указывающими на вероятность активности белка в соответствующих локусах, особенно если они попали, например, в гены, кодирующие жизненно важные белки. Это будет означать, что такую конфигурацию не стоит далее рассматривать даже *in silico* из-за высокого риска проявления побочных эффектов.

Процесс сканирования всего генома может занять длительное время. Поэтому остро стоит вопрос об ускорении вычислительных процессов, так как разработка безопасных искусственных белков предполагает, вообще говоря, проверку не двух-трёх конфигураций, а перебора нескольких десятков, а лучше, даже сотен и тысяч различных вариантов, для выбора самого безопасного из них. Ввиду проблем связанных с длительной задержкой результатов, некоторые разработчики даже пошли на ухищрения, вспомнив что у многих организмов, включая человека, до 95% и выше от всей протяжённости генома приходится на так называемую мусорную ДНК, и считается, что изменения нуклеотидного состава, попытки транскрипции, какая-нибудь другая активность в этих местах не могут привести к пагубным последствиям, поэтому они проверяют совпадения только в 3–5 % от ДНК, таким образом многократно ускоряя выдачу результатов. Более того, если новый TALE разрабатывается как искусственной фактор транскрипции, для ещё большего ускорения объём поиска можно сократить до списка известных промоторов всех генов [28].

Мы, имея некоторый опыт в деле оптимизации и ускорения задач, связанных с распознаванием визуальных образов, решили использовать такие же подходы для ускорения задач, связанных с поиском в генетических последовательностях. Первое, что следует сделать, это освободить сводную таблицу от незначительных величин, обнулив их ячейки, что в дальнейшем позволит упростить все вычисления. Например, если применять алгоритм Брэдли для адаптивного определения порогового значения [29], то следует сначала определять среднее значение для каждой колонки, назначая им в качестве порогового значения величину на 5 %, ниже среднего, и обнулять все ячейки, содержимое которых не достигает этой величины. Причём среднее значение, очевидно, вычисляется так:

$$\bar{P}_{\text{RVD}} = \frac{1}{4}\sum\nolimits_n P_{\text{RVD:}n},\ n \in \{A,C,G,T\}.$$

Оставшиеся ячейки будем округлять до круглых чисел, причём малые значения (<100) будем округлять более деликатно, разбивая диапазон значений с шагом равным пяти, а выше с шагом в десять раз большим, т. е. применим *неравномерное квантование* вида:

$$P'_{\text{RVD:Nuc}} = \begin{cases} 0, & P_{\text{RVD:Nuc}} < \frac{1}{5}\sum\nolimits_n P_{\text{RVD:}n}; \\ \text{round}\left(P_{\text{RVD:Nuc}}, 5\right), & P_{\text{RVD:Nuc}} < 100; \\ \text{round}\left(P_{\text{RVD:Nuc}}, 50\right), & P_{\text{RVD:Nuc}} \geq 100. \end{cases}$$

где P – количество наблюдаемых в природе соответствий данного типа точечной связи RVD:нуклеотид, P’ – это новое квантованное значение этой величины, $n \in \{A,C,G,T\}$. Полученные квантованные значения представлены в таблице 2.

Таблица 2. Значения квантованных значений для вероятностей связей RVD:нуклеотид

Нуклеотиды	NI	HD	NN	NG	NS	N-	HG	H-	IG
A	100	0	15	0	45	0	0	0	0
C	0	150	0	0	20	30	0	0	0
G	0	0	35	0	0	0	0	0	0
T	0	0	0	100	0	10	15	0	0
G-mean (+1)	3.17	3.51	4.9	3.17	5.57	4.3	2	1	1

Теперь всё готово к применению логарифма отношения шансов, за исключением того момента, что в случае нулевых значений логарифм отношения обращается в минус бесконечность – эта проблема с сингулярностью часто решается технически, когда все представленные значения перед вычислением логарифма будут увеличены на единицу (инкрементированы). Второй момент связан с выбором масштабного коэффициента, потому что ускорение вычислений, связанных с матрицами оценок, при реализации на компьютере предполагает переход к целочисленной арифметике, что приводит к необходимости округления полученных оценок после умножения на некий коэффициент. Например в случае таблиц BLOSUM он выбран равным двум, а здесь коэффициент подбирается из расчёта, что четыре последовательных балла никогда не выйдут за границы диапазона значений (–127,+127). Этого требует специфика нашей реализации, т. к. дополнительное ускорение при расчёте суммарных оценок достигается на процессорном уровне за счёт объединения соседних четвёрок элементов в отдельные тетрады, для которых значения суммы их оценок заданы в виде таблиц поиска – таким образом используя инструкцию XLAT для классического ассемблера x86 удалось вчетверо ускорить вычисления ещё до применения SIMD инструкций. Учитывая всё вышесказанное, приходим к формулам:

$$\begin{cases} s_{\mathrm{RVD:Nuc}} = K \cdot \ln \dfrac{P'_{\mathrm{RVD}:n} + 1}{\sqrt[4]{\prod_n \left(P'_{\mathrm{RVD}:n} + 1 \right)}}, \\ \max\limits_{\mathrm{RVD}} \left(\max\limits_{n} \left| s_{\mathrm{RVD}:n} \right| \right) = \dfrac{127}{4}, n \in \{A, C, G, T\}, \mathrm{RVD} \in \{\mathrm{NI}, \mathrm{HD}, \ldots\}. \end{cases}$$

где K – коэффициент масштабирования, задаваемый вторым условием данной системы. Для рассматриваемых в данной работе данных коэффициент равен $K \approx 8.4$, и тогда все индивидуальные оценки для связей RVD-нуклеотид представлены в таблице 3.

Таблица 3. Итоговые значения для оценки точечных связей RVD:нуклеотид

Нуклеотиды	NI	HD	NN	NG	NS	N-	HG	H-	IG
A	29.21	–10.58	9.99	–9.74	17.81	–12.30	–5.85	0	0
C	–9.74	31.75	–13.41	–9.74	11.19	16.67	–5.85	0	0
G	–9.74	–10.58	16.83	–9.74	–14.50	–12.30	–5.85	0	0
T	–9.74	–10.58	–13.41	29.21	–14.50	7.93	17.55	0	0
K =	8.437506918								

Предсказание простого (непарного) TALE-ДНК связывания

При разработке искусственных TALE различают два принципиально разных случая: простое TALE-ДНК связывание (рис.5) и парное. Последний из них в природе не встречается, а вот непарный случай действительно прост. Пусть требуется оценить связь TALE белка с некоторым числом L повторов и текущим ДНК участком (той же длины L):

$$\left(\mathrm{TALE} = \mathrm{RVD}_1, \mathrm{RVD}_2, \ldots, \mathrm{RVD}_L\right) : \left(\mathrm{site} = \mathrm{Nuc}_1, \mathrm{Nuc}_2, \ldots, \mathrm{Nuc}_L\right),$$

тогда несложно предложить суммарную оценку вероятности такой связи в виде:

$$S_{\text{TALE:site}} = \sum_{i=1}^{L} s_{\text{RVD}_i:\text{Nuc}_i},$$

где единичные оценки s берутся из таблицы 3.

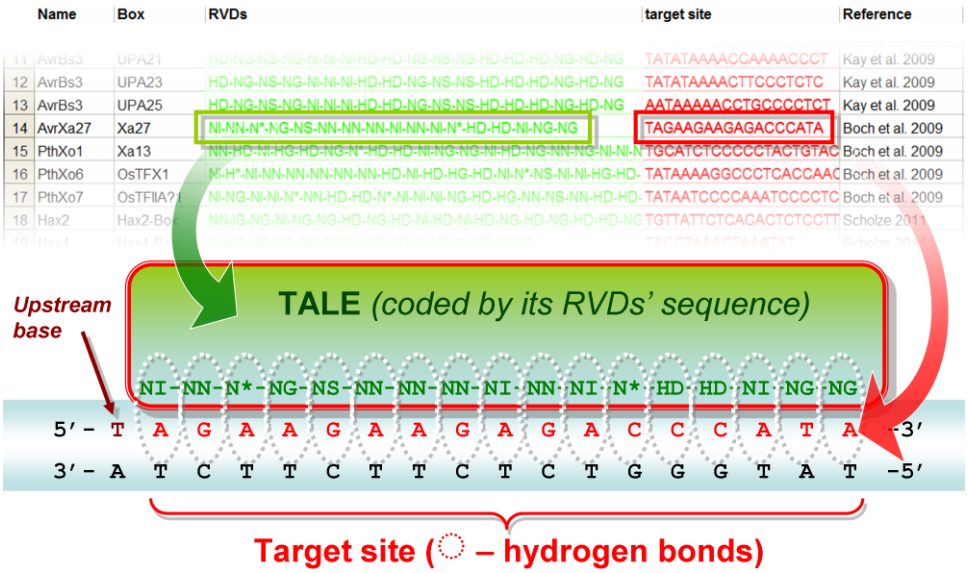

Рис. 5. Схематическое представление связывания группы из RVD остатков и участка двойной нити ДНК посредством водородных связей.

Далее потребуется предложить некоторую простую проверку для выделения всех потенциальных участков связи из общего ряда всех рассматриваемых участков на анализируемой ДНК. Наиболее простым из проверок является введение некоторого порогового значения θ такого, что все участки, суммарные оценки на которых оказались выше этого порога, будут попадать в результирующий список вероятных участков связи, в то время как все остальные кандидаты будут просто игнорироваться. Другими словами, для выборки всех участков ДНК длины L определяется фильтр с одним простым условием:

$$S_{\text{TALE:site}} \geq \theta_{\text{TALE}}.$$

Число повторов L для различных белков может варьироваться, т. е. задание универсального абсолютного порогового значения для всех TALE белков, в принципе, невозможно, т. к. порог для суммы из большего числа связей, очевидно, должен быть выше, чем из меньшего. Поэтому пороговое значение принято задавать не в абсолютных значениях, а относительно, как некоторый процент от лучшей, максимально возможной суммарной оценки, которая может быть получена для данного конкретного TALE белка (для его RVD последовательности):

$$\theta_{\text{TALE}} = k \sum_{\text{RVD}=\text{RVD}_1}^{\text{RVD}_L} \max_n s_{\text{RVD}:n}, n \in \{A, C, G, T\}.$$

где k – задаваемый пользователем коэффициент фильтра (обычно k = 0.6). Таким образом, предложена общая вычислительная модель в случае предсказания простого TALE-ДНК связывания. Результаты предсказания, получаемые на основе простых моделей с суммирующими оценками, можно считать вполне удовлетворительными, но они требуют существенного повышения качества предсказания. Для улучшения модели предсказания в онлайн-сервисе TALENoffer [20] применён так называемый «контекстный» принцип расчёта оценок, когда вероятность успешной связи RVD-нуклеотид рассчитывается на основе анализа всех RVD и нуклеотидов в соседних позициях, что дало явное улучшение предсказания, но привело к резкому увеличению времени вычислений.

Заметим, что до сих пор обсуждалась вероятность связи TALE белка лишь с верхней цепью ДНК, т. е. в направлении 3'-конца, но следует помнить, что TALE белок способен

связаться и с нижней цепью ДНК (но в направлении к 5'-концу) – поэтому при организации соответствующих вычислений следует дважды проверять (сканировать) участки ДНК, как для прямой, так и для комплементарной (инвертированной) цепи.

В данной работе удалось предложить вычислительную модель, результаты предсказания которой сравнимы по качеству с более сложной из общепризнанных моделей [20], но столь же высокопроизводительную, как [19]. Этого удалось добиться за счёт введения нескольких этапов анализа участков-кандидатов, но не по ходу сканирования ДНК, а потом, на стадии постобработки результатов, полученных простой моделью. Ниже изложены принципы работы всех трёх наших этапов постобработки, благодаря чему удалось значительно улучшить все основные показатели предсказания.

Постобработка I: Отсев участков с сайтами метилирования

Обычно списки потенциальных участков связи при любом *in silico* предсказании приносят гораздо больше предполагаемых побочных сайтов связывания, чем это возможно проверить в ходе реальных *in vitro* экспериментов. Точно так и вывод потенциальных сайтов связывания TALE-ДНК даже для одной нити ДНК может составлять сотни и тысячи элементов и даже больше (ведь при снижении порогового значения θ требования к кандидатам также снижаются и предварительный список пополняется ещё большим числом участков, требующим дополнительной проверки). Однако существует возможность значительного сокращения списка предварительных результатов без опускания порога, если исключать из него те участки ДНК, которые заведомо не могут быть связаны с данным белком согласно некоторым дополнительным критериям отбора. Такие критерии представляются обычно в виде опций поиска, но в рамках нашего инструмента они рассматриваются как этапы постобработки этого списка.

Изложим первый из критериев, успешно применённый в рамках данной работы. Так, в ряде работ делалось предположение [30], почему некоторые участки с высокой расчётной вероятностью связывания не проявляют такой активности – видимо это потому, что они содержат динуклеотидные сайты метилирования [31]. А значит, согласовывая результаты предварительного списка с картами возможного метилирования для рассматриваемого ДНК, и исключив из списка все такие элементы с участками метилирования, можно значительно сократить список. Заметим, что на стадии сканирования ДНК подобная проверка отняла бы много времени, но как этап постобработки даже подробный анализ сотен и тысяч элементов предварительного списка займёт гораздо меньше времени, чем составление этого списка.

Вспомним, что необходимым условием метилирования участка является присутствие пары нуклеотидов цитозина и гуанина, расположенных в таком порядке CG (так называемый CpG-островок), однако наличие CG-динуклеотида не является достаточным условием метилирования и поэтому существует ряд методов для составления точных карт метилирования различных участков ДНК, например, с помощью метилировано-специфичной ПЦР [32]. Показано, что от 70 % до 80 % всех CG-динуклеотидов в ДНК млекопитающих подвержены метилированию, что позволило нам упростить этот критерий постобработки просто предложив исключать из рассмотрения любой потенциальный участок TALE-ДНК связывания, если в нём обнаружен хотя бы один CG-динуклеотид (а значит существует высокая вероятность метилирования). Заметим, что даже в подобной радикальной постановке данный критерий хорошо показал себя на тестовых выборках.

Постобработка II: Отсев участков с «плохим затактом»

Вернёмся к оценке вероятности связи между повторами некоторого TALE белка и текущим ДНК участком длины L, обратив внимание на то, что индекс позиции i в формуле меняется от 1 до L. Очевидно, что следующие позиции в направлении 3'-конца (L+1, L+2) никак не повлияют на суммарную оценку S, но это неверно для первой позиции перед самим участком (соответствующей i=0). Нуклеотид, находящийся в этой позиции по

отношению к рассматриваемому участку, в англоязычной литературе часто обозначается как «upstream base pairs/nucleotide», мы для краткости будем называть его «затактом».

Ещё в работе [19] обнаружено, что для всех известных на тот момент случаев связывания натуральных TALE и участков ДНК растений в позиции «затакта» всегда находился нуклеотид T (тимин), и поэтому самые первые реализации инструментов предсказания TALE-ДНК игнорировали любые другие участки (без T в позиции i=0). Позже, с накоплением обнаруживаемых в природе случаев связывания, оказалось, что C (цитозин) в этой позиции также может быть приемлем (хотя встречается реже). Ещё позже стал известен ещё более редкий, но всё ещё возможный случай связывания для участков с затактом A (аденином). Наконец, в ходе *in vitro* экспериментов, показано, что иногда (для участков с высокой суммарной оценкой) не представляет сложности связаться с участком, имеющем G (гуанин) в нулевой позиции. Значит можно к суммарной оценке добавить ещё и баллы за хороший затакт:

$$\text{score\_0} = \{T{:}+30,\ C{:}+10,\ A{:}-10,\ G{:}-30\},$$

вытесняя многие участки с плохим затактом на последние позиции, где они отфильтровываются пороговым значением. Заметим, что данный критерий хорошо показал себя на имеющихся тестовых выборках, поэтому был введён авторами как дополнительный этап постобработки для предварительных списков предсказания и оценки участков связывания TALE-ДНК.

Постобработка III: Отсев «незастёгивающихся» участков

Следующий, заключительный этап постобработки задаётся ещё одним критерием отсева участков-кандидатов из предварительного списка, который более сложен, чем предыдущие два, но даёт наибольшее улучшение качества предсказания. Этот критерий основан на некотором гипотетическом представлении о физическом характере взаимодействия центрального домена TALE и участка ДНК в момент их вероятного связывания. Согласно этим представлениям, вводится новая математическая модель для проведения на заключительном этапе более точной оценки вероятности связи TALE с каждым конкретным участком из предварительного списка. Введение этого этапа постобработки, с одной стороны, позволяет получить принципиально более точное предсказание в сравнении с другими подобными инструментами, использующими лишь простую модель суммарной оценки (как реализовано в работе [19]). А с другой стороны, этот этап практически никак не отражается на времени вычислений, что выгодно отличает его от сложных моделей, как в [20], которые дают предсказание, сравнимое по качеству с нашим, но при этом являются неоправданно медлительными.

Изложим положения, касающиеся физического представления о взаимодействии TALE-повторов из центрального домена белка с участком ДНК. При сравнительном анализе участков связывания, обнаруженных *in vitro* и предсказанных для этого же случая *in silico*, легко заметить наличие существенных расхождений, когда участок связывания оказывался недооценённым или, наоборот, переоценённым, но никак не проявил себя на практике. При изучении таких данных мы хотели выявить некоторое простое правило, которое позволило бы пересортировать предварительный список предсказания так, чтобы все недооценённые участки поднялись в списке выше, а переоценённые ниже.

Дальнейший сравнительный анализ подтвердил догадку, выдвинутую в работе [33], о том, что на успешность связывания TALE с ДНК участком влияет не только общее число неудачных совпадений RVD-нуклеотид, и не только общая сумма штрафов за все эти несовпадения, но и их взаимное расположение. Но в отличии от работы [20], мы стремились к выявлению простого, физически обоснованного правила, хотя бы частично оправдывающего такое расхождение, предпочитая этот подход построению сложной математической модели на основе марковской модели, имеющей десятки параметров, подбор и коррекция которых осуществлялась в ходе обучения нейронной сети [20].

Такое простое правило успешно найдено после сравнения участков неудачных

точечных RVD-нуклеотид связей и обнаружения интересной закономерности: даже меньшее число точечных неудач оказывается фатальным, если они расположены в ряд, друг за другом, в то время как даже большее число «плохих» точек связи не станет критичным, если все неудачные точечные связи разбросаны, а не скучены в одном месте.

Центральный домен TALE белка с помощью ряда водородных связей пытается «пристегнуться» к почти идеально подходящему для него участку ДНК, но неожиданно (с точки зрения простой модели предсказания) этой связи не наблюдается *in vitro*. Получается, что на последнем этапе предварительный список должен быть пересортирован так, чтобы участки, к примеру, с шестью «плохими» точечными связями, но разбитыми на группы, оказались в списке выше участков с пятью «плохими» точками, если те скучены в одном месте (рис. 6). Так приходим к простому правилу, которое неформально можно записать в виде: все участки TALE-ДНК связывания, имеющие достаточно протяжённые места плохого связывания, должны быть исключены из предварительного списка кандидатов, даже несмотря на их высокую суммарную оценку.

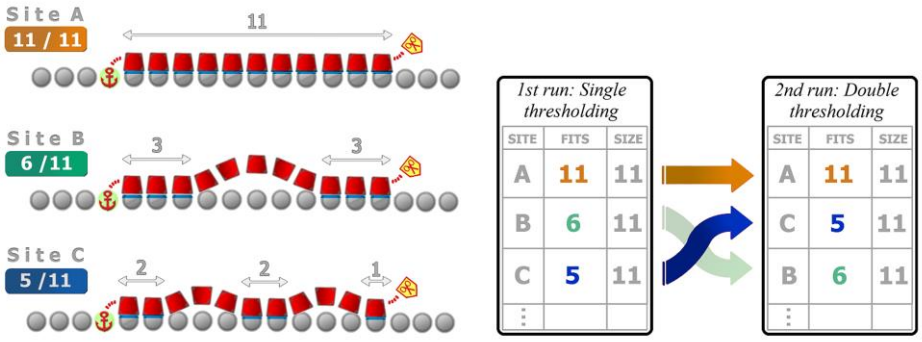

Рис. 6. Схематическое представление связывания TALE:ДНК, наглядно демонстрирующее случай, когда связывание с лучшими суммарными оценками может уступать связыванию с меньшей расчётной оценкой, но с распределёнными ошибками (модель «сломанная молния»).

Теперь перейдём к математической формализации этого правила, предложив далее конкретную вычислительную реализацию. Очевидно, что рост длины «плохой» группы точечных связей в таком случае должен приводить к нарастанию некоторой массы ошибок и при наступлении некоторой «критической массы» вероятность связи с данным участком ДНК становится нулевой, вне зависимости от того, насколько идеальным окажутся следующие точечные связи RVD-нуклеотид. Пусть, к примеру, «плохой» точечной связью будет считаться любая связь RVD-нуклеотид с отрицательным значением из таблицы оценок и пусть за массу ошибок принят размер непрерывной «плохой» группы – B, тогда самым простым численным решением для такого правила может служить простое условие фильтра: $B > X$, где X – некоторая заранее определённая пороговая величина для максимально возможной «плохой» группы. Интересно, что даже в этой грубой постановке введённое нами правило позволило улучшить качество предсказания на тестовых выборках, но эта первая реализация оказалось слишком примитивной и грубой для того, чтобы учесть, насколько «плохи» отдельные плохие группы одинаковой длины.

Более деликатный способ расчёта «критических показателей» для связей TALE-ДНК можно предложить на основе ряда уравнений свёртки, учитывающих эридитарную («наследственную») связь в рассматриваемых последовательностях. После тестовых проверок авторы остановили выбор на применении экспоненциальных скользящих средних для преобразования ряда точечных оценок и выявления критических мест связывания. Таким образом получаются новые, экспоненциально сглаженные значения $\{s'\}$, которые и служат оценкой «критичности» в данной конкретной позиции. При этом критичность в нулевой позиции могла бы и зависеть от «затакта», но для простоты здесь s'_0 принимается равным нулю. Остальные члены ряда оценки критичности рассчитываются согласно рекуррентной формуле:

$$\begin{cases} s_{\mathrm{RVD}_0:\mathrm{Nuc}_0} = 0 \\ s'_{\mathrm{RVD}_i:\mathrm{Nuc}_i} = \alpha s_{\mathrm{RVD}_i:\mathrm{Nuc}_i} + (1-\alpha)\, s'_{\mathrm{RVD}_{i-1}:\mathrm{Nuc}_{i-1}} \end{cases},$$

где α = ½ – сглаживающая константа, подобранная опытным путём. Теперь остаётся лишь ввести столь же простое условие для фильтра:

$$\left\{\mathrm{site} \in \mathrm{output} \vee \forall s'_{\mathrm{RVD}_i:\mathrm{Nuc}_i} > \lambda\right\},$$

где λ = –10 – локальное пороговое значение, подобранное опытным путём на основании анализа тестовой выборки. В рамках реализации нашего инструмента решено предоставить пользователю возможность свободного изменения данного параметра, что позволяет им значительно сокращать результирующий список участков-кандидатов побочного TALE-ДНК связывания. Очевидно, если слишком занизить локальное пороговое значение (например, допустить λ = –100), то список участков-кандидатов на этапе постобработки вообще никак не изменится, а вот при высоком значении (λ = 0) рискуем потерять почти всех таких кандидатов.

Предсказание парных TALE-ДНК связываний

Сравнительно недавно появился инструмент геномного редактирования, в котором основная роль ДНК-ножниц отводится сразу паре специально разработанных искусственных TALEN, которые получаются из обычных TALE путём замены их последнего, третьего домена (отвечавшего за принудительную активацию транскрипции в организме-хозяине) на нуклеазу (обычно Fok1). Как и простые TALE, искусственные нуклеазы TALEN могут быть перепрограммированы и нацелены на любой участок цепи ДНК, а значит, в случае связывания TALEN-ДНК, сразу за этим участком (в направлении 3'-конца) произойдёт разрыв данной цепи ДНК.

Для задач геномного редактирования недостаточно одиночного разрыва двойной спирали ДНК, поэтому одновременно с нацеливанием на участок первой (верхней) цепи ДНК исследователи разрабатывают парный TALEN, который метит приблизительно в тоже место, но в комплементарной цепи ДНК. Только в случае успешной посадки обеих TALEN, обрамляющих место предполагаемых одиночных разрывов, можно получить двойной разрыв ДНК – таким образом пара TALEN действует сообща, как лезвия ножниц (рис. 7).

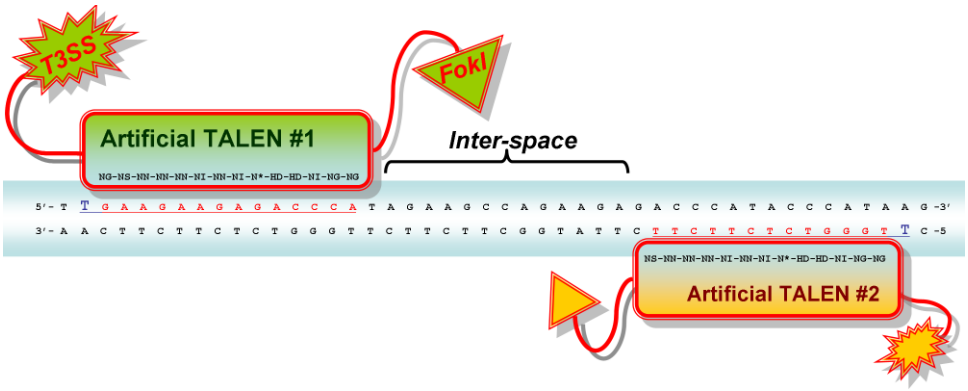

Рис. 7. Схематическое представление двойного разрыва с помощью пары TALEN.

Практическое использование любой новой синтезированной пары TALEN без предварительного предсказания побочного поведения этой пары крайне нежелательно, т. к. может привести к непредсказуемым побочным изменениям во всех местах генома, где находится похожая пара участков ДНК с похожим нуклеотидным составом и на достаточно близкой дистанции друг от друга (что ведёт к опасности появления нежелательного разрыва и редактирования ДНК). Чтобы избежать побочных модификаций генома, следует заранее предсказать все такие возможные ситуации и оценить вероятность

каждого из них, на основании чего и следует принимать окончательное решение о синтезе любой пары TALEN, или же, в случае высокого риска, попробовать подобрать другие, менее рискованные целевые участки.

Таким образом приходим к необходимости разработки, по сути, нового инструмента для предсказания парного связывания, который может быть построен на основе моделей, изложенных выше. Совместная вероятность двух независимых событий равна произведению их вероятностей. Однако при рассмотрении на логарифмической шкале оценка вероятности для связывания двух TALE равна не произведению, а сумме их независимых оценок, т. е.:

$$J_{1,2} = S_{\mathrm{TALE}_1:\mathrm{site}_1} + S_{\mathrm{TALE}_2:\mathrm{site}_2}.$$

Тогда, если обозначить совместное пороговое значение, к примеру, как:

$$\Theta_{1,2} = 2\max\left(\theta_{\mathrm{TALE}_1}, \theta_{\mathrm{TALE}_2}\right),$$

то для случая парного связывания перепишем условие фильтра в виде:

$$J_{1,2} = S_{\mathrm{TALE}_1:\mathrm{site}_1} + S_{\mathrm{TALE}_2:\mathrm{site}_2} \Theta_{1,2}.$$

Несмотря на то, что полученное условие отсева кандидатов является обоснованным и логичным, некоторые разработчики [19] используют другое, статистически менее корректное условие фильтра:

$$\left(S_{\mathrm{TALE}_1:\mathrm{site}_1} \geq \theta_{\mathrm{TALE}_1}\right) \cap \left(S_{\mathrm{TALE}_2:\mathrm{site}_2} \theta_{\mathrm{TALE}_2}\right),$$

т. к. оно позволяет сильно упростить предсказание попарного связывания для двух TALE сведя его к нахождению хороших связей по отдельности для каждого из них, а потом совмещению этих списков – но таким образом они упускают случаи вида «отличный левый участок / плохой правый», которые на практике в разы эффективнее, чем «хороший левый / хороший правый».

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Исторически TALEN появились как лучший аналог технологии ZFN, но с открытием CRISPR интерес к ним временно ослаб. Казалось, что дешевизна и большое разнообразие нативных эффекторных ферментов для конструирования CRISPR позволят побороть активность на нецелевых участках ДНК. Позже выяснилось, что все разновидности CRISPR-систем часто проявляют нецелевую активность [1], что сильно снижает надежность инструмента и усложняет предсказание потенциальных сайтов связывания.

К сожалению, никто так и не провел до сих пор сравнения частоты нецелевой активности TALEN и CRISPR. Поэтому строгих доказательств в пользу б**о**льшей точности TALEN пока нет, есть только несистематизированные данные. Но бесспорны факты в пользу TALEN, что технология за последнее время подешевела; и что CRISPR не может нацеливаться на любой участок генома, потому что направляющая РНК требует наличия консервативного протоспейсерного мотива (PAM) рядом с участком связывания. Тем не менее, процесс создания TALEN требует дорогостоящих *in vitro* экспериментов на основе SELEX. В настоящее время многие исследователи полагают, что в скором времени всё же станет возможным заменить дорогостоящие и длительные предсказания *in vitro* чисто компьютерным моделированием.

Таким образом, в последнее время работы по развитию технологии TALEN снова становятся актуальны. TALEN являются перспективным инструментом биоинженерии с широким спектром применений, начиная с целенаправленной модификации генома сельскохозяйственных культур, редактирования геномов модельных организмов с целью их дальнейшего изучения, и заканчивая развитием методов лечения различных негенетических заболеваний.

Существующие в настоящее время веб-сайты и инструменты помогают ученым правильно позиционировать и разрабатывать TALEN. Многие из этих инструментов

выполняют множество различных функций, но их основная цель — выявить возможные сайты связывания TALEN в конкретной последовательности ДНК, выбрать подходящие RVD и, в некоторых случаях, предсказать нежелательные отклонения от цели. Вот некоторые из них: TALE-NT [19], idTALE [34], PROGNOS [35], TALENoffer [20], E-TALEN [36], Mojo Hand [37] и ChopChop [38].

Мы провели сравнительный анализ качества предсказания нашей программы TANDIS с TALE-NT [19] и web-серверной реализацией на платформе GALAXY алгоритма TALENoffer [20]. Результаты сравнительных исследования представлены на рисунке 7. Предсказание *in silico* мы проводили для четырех генов, для которых есть данные предсказаний *in-vitro* методом SELEX, [17, 18], считающиеся непререкаемым эталоном для *in-silico* предсказаний. Наш метод оказался достаточно эффективным по качеству предсказания. Результаты TANDIS сравнимы с Galaxy/TALENoffer [20], но наша модель сломанной молнии намного проще (оперирует всего 4 параметрами), поэтому сканирование генома проходит в 2–2.5 раза быстрее, чем с помощью онлайн-сервиса Galaxy/TALENoffer.

Перед началом сравнения мы сортируем все выходные результаты по суммарным «очкам» (т. е. сумма первой и второй оценки связывания). Так, можно утверждать, что один подход лучше другого, если все участки, что подтверждены *in-vitro*, попадают в самый верх его списка результатов, в то время как в списке конкурента эти сайты разбросаны по всему списку, либо вовсе отсутствуют (что совсем плохой результат, помечаемый нами N/A). В таблице 4 приведены координаты целевых сайтов связывания, подтверждённые *in-vitro* и соответствующие им числовые ранги в списках результатов, полученных тремя различными *in-silico* инструментами.

Естественно, что для всех инструментов использованы одни и те же настройки: мы ищем все возможные комбинации левый/правый сайт и диапазон для дистанции, установлены аналогично (12–24 пар оснований). Ни Galaxy, ни наш инструмент TANDIS по умолчанию не имеют ограничений в отношении базовой upstream пары, но в TALE-NT это учитывается, поэтому устанавливаем самый общий вариант (T или C), чтобы уравнять наши шансы найти все участки. Кроме того, все пороговые значения используем по умолчанию и для Galaxy (фильтр $q = 0{,}4$), и для TANDIS (глобальный порог = 60%), но для TALE-NT используем параметр самого низкого возможного порога (значения отсечки 4.0) – увеличиваем его шансы найти все сайты. Остальные специфические параметры TANDIS брались по умолчанию:

1) все метилированные сайты (содержащие любой подсайт C-G) должны быть исключены из выходного списка;

2) не используются бонусы для хороших затактов;

3) локальный порог установлен на значение по умолчанию (–10).

Как видно из таблицы 4, TANDIS и Galaxy получали лучшие значения, причём TANDIS оказывался лучшим 11 раз, а Galaxy 6 раз, оставляя далеко позади TALE-NT. Также интересно сравнить время работы сервисов. Все три сервиса продемонстрировали корреляцию между числом TALE повторов и временем выполнения задачи, что вполне ожидаемо. Так, для 15–18 RVD время работы изменяется для TALE-NT от 35 до 120 минут, для Galaxy от 25 до 40 минут и от 5 до 7 минут для TANDIS.

В качестве обучающей выборки для сервиса TANDIS выбраны результаты *in-vitro* предсказания при таргетировании гена PPP1R работы [17], а тесты проведены на результатах таргетирования генов COL7A1, CCR5R, ATM из [18]. Причём, RVD последовательности для всех сервисов *in silico* предсказания берутся в точности, как они приводятся в экспериментальных работах для каждого из участков таргетирования методом SELEX. Считается, что частота использования отдельных RVD способна влиять на «общую успешность» связывания всей искусственной TALE, поэтому ниже приведены также частоты встречаемости RVD в каждой из рассматриваемых TALE последовательностей.

Отметим, что сравнение нашего подхода с популярными сервисами показало, что ни один из подходов не дает идеального решения, у каждого инструмента есть сильные и слабые стороны.

Таблица 4. Прямая верификация результатов предсказания TANDIS и сравнение с результатами предсказания сервисами TALE-NT [19] и Galaxy/TALENoffer [20]. Приведена в сравнении с результатами четырёх *in-vitro* предсказаний на основе SELEX [17, 18]

***in vitro* prediction**					***in silico* prediction**		
Site	**Sequence**	**Left 5'-position**	**Inter-spacer**	**Right 5'-position**	**TALE-NT**	**Galaxy**	**TANDIS**
PPP1R12C [1]	Chr. 19	55627107	15	55627157	1	1	1
OTS 10	Chr. 15	67305690	24	67305738	N/A	**197**	1171
OTS 14	Chr. 5	165831362	16	165831421	N/A	**48**	65
COL7A1 [2]	Chr. 3	48628076	24	48628131	1	1	1
PRMT6	Chr. 1	106271407	24	106271462	N/A	N/A	N/A
PRMT2	Chr. 21	48094457	17	48094505	N/A	3	**2**
GGT1	Chr. 22	24991120	17	24991168	76	**106**	563
CCR5A [3]	Chr. 3	46414913	18	46414968	N/A	1	1
OffC-5	Chr. 2	49729850	14	49729901	N/A	5732	**4189**
OffC-15	Chr. 4	167658880	26	167658943	N/A	N/A	N/A
OffC-16	Chr. 9	37603268	28	37603333	N/A	N/A	N/A
OffC-28	Chr. 12	1809633	19	1809689	18	**3**	49
OffC-36	Chr. 17	72377593	21	72377651	44	**112**	1287
OffC-38	Chr. 16	10752185	16	10752238	N/A	920	**314**
OffC-49	Chr. 7	113045477	15	113045529	N/A	2727	**43**
OffC-69	Chr. 6	143986785	23	143986845	N/A	972	**733**
OffC-76	Chr. 3	189629641	19	189629697	N/A	**198**	262
OnATM [3]	Chr. 11	108098593	18	108098648	N/A	1	**1**
OffA-1	Chr. 16	26685904	20	26685961	N/A	140	**28**
OffA-11	Chr. 5	80235624	16	80235677	N/A	7148	**3300**
OffA-13	Chr. 3	174452887	17	174452941	N/A	8808	**1136**
OffA-16	Chr. 16	56199461	24	56199522	N/A	1988	**72**
OffA-17	Chr. 4	161404932	15	161404984	N/A	N/A	**14485**
OffA-23	Chr. X	86852027	15	86852079	N/A	331	**89**
OffA-35	Chr. 1	77009092	29	77009158	N/A	N/A	N/A

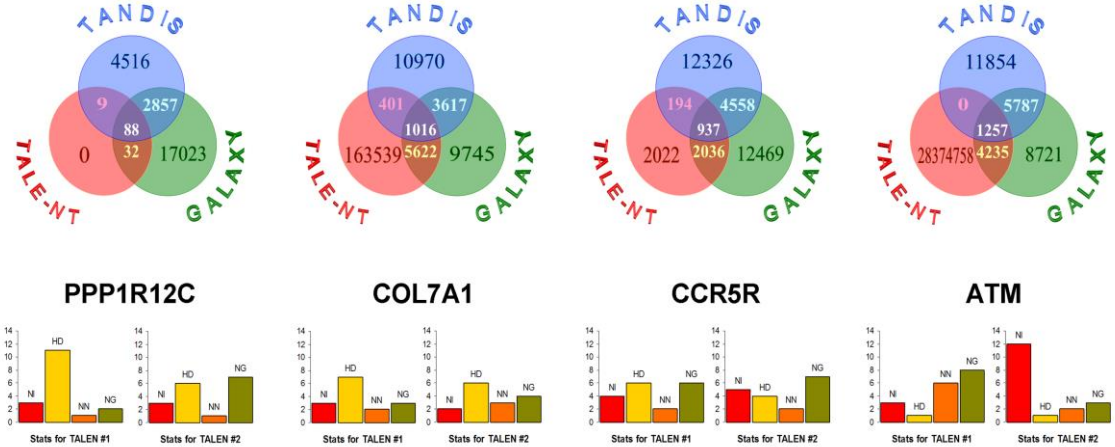

Рис. 8. Косвенная верификации и сравнительный анализ результатов предсказания тремя различными инструментами TALE-NT, GALAXY/TALENoffer и TANDIS.

ЗАКЛЮЧЕНИЕ

Несмотря на многообещающие перспективы CRISPR-Cas систем, технологии редактирования геномов, использующие ферменты, способные распознавать и расщеплять ДНК (ZFN и TALEN) не исчезают из поля зрения биологов. Эти технологии оптимизированы для разрезания целевых последовательностей ДНК и внесения мутаций в различных приложениях. Чтобы редактирование генома было применимо в медицине, восстановление разрезанной ДНК должно быть полным, а восстановленный геном не должен содержать ошибок. Однако этих целей в настоящее время трудно достичь с помощью внутреннего механизма репарации ДНК, поскольку путь репарации ДНК сложен и неуправляем.

Альтернативный подход к регулированию функции генов заключается в модификации геномной ДНК для контроля экспрессии генов без изменений или с минимальными изменениями в базовой ДНК. В этом случае речь идет о редактировании эпигенома. Такая методика считается потенциальным терапевтическим подходом к различным генетическим заболеваниям и некоторым видам рака. Эти генетические заболевания, в основном, затрагивают гены с усилением функции, поскольку генетические заболевания с потерей функции можно лечить с помощью традиционной генной терапии [39]. Редактирование эпигенома подходит и для лечения ненаследственных заболеваний, поскольку экспрессию целевого гена можно увеличить или уменьшить путем модификации целевого гена [40]. Более того, редактирование эпигенома представляет собой практический подход к искусственному манипулированию структурой хроматина произвольных областей генома.

Все клетки человеческого тела имеют одинаковую генетическую информацию. Во время дифференцировки клетки изменения в структуре ее хроматина обеспечивают активную транскрипцию нужных генов, а ненужные гены не транскрибируются из-за образования гетерохроматиновых структур [41]. Эти структуры и фенотипы сохраняются после каждого клеточного деления. Эпигенетические процессы в клетке тесно связаны с химическими модификациями ДНК, гистонов, белков хроматина и некодирующих РНК [42, 43]. За исключением рекомбинации ДНК в иммунных клетках, последовательность геномной ДНК большинства типов клеток сохраняется независимо от судьбы дифференцировки. В частности, при ацетилировании хвостов гистонов структуры хроматина становятся расслабленными и доступными для РНК-полимераз и основных транскрипционных факторов, которые активируют транскрипцию гена [44]. С другой стороны, когда ДНК метилируется, или хвосты гистонов подвергаются репрессивным посттрансляционным модификациям, хроматин конденсируется и становится недоступным для РНК-полимераз и факторов транскрипции; таким образом, экспрессия генов подавляется [41].

Системы CRISPR-Cas, использующие направляющие РНК, лежат в основе методов генной терапии, потому что они направлены на модификацию геномной ДНК. Технологии на основе цинковых пальцев или TALE могут нести в качестве ассоциированных эффекторов ферменты, способные модифицировать эпигеном, а не геномную ДНК. Поэтому их можно применять в качестве белковых препаратов аналогично доступным коммерческим лекарствам [45]. Поскольку иммуногенность белковых лекарственных препаратов изучается более широко, чем генная терапия, доступность цинковых пальцев и TALE в качестве белковых лекарств может быть основным преимуществом при разработке эпигеном-модифицирующих ферментов в качестве терапевтических лекарств [46].

Основная особенность редактирования эпигенома, заключающаяся в сохранении нуклеотидной последовательности, считается одновременно слабой и сильной стороной, поскольку ген, вызывающий заболевание, не изменяется. Вместо этого редактирование эпигенома подавляет экспрессию генов, вызывающих заболевания, или увеличивает экспрессию генов, таких как гены клеточного цикла и другие подавленные гены.

Редактирование эпигенома может привести к меньшему количеству побочных эффектов, поскольку оно нацелено только на одну или несколько последовательностей в геноме и не действует на ферменты, модифицирующие эпигеном. Более того, модифицированные структуры хроматина сохраняются после деления клеток, поскольку они используют эндогенный механизм эпигенетического поддержания внутри клеток. Еще одним преимуществом редактирования эпигенома является его обратимый эффект по сравнению с необратимыми изменениями последовательности ДНК при редактировании генома.

Ферменты, модифицирующие эпигеном, такие как ацетилтрансферазы гистонов, ДНК-метилтрансферазы и транслокационная метилцитозиндиоксигеназа 1 (TET1), которая представляет собой метилцитозиндиоксигеназу, деметилирующую метилированную ДНК, играют важную роль в изменении эпигенома. Однако, у большинства ферментов, модифицирующих эпигеном, нет ДНК-связывающего домена, который определяет целевую последовательность. Их целевая специфичность зависит от типа транскрипционных факторов, с которыми они образуют комплексы [47].

Факторы транскрипции связываются с несколькими генами-мишенями (часто от сотен до тысяч), а длина целевой последовательности отдельного фактора транскрипции колеблется от нескольких до десятков оснований. Поэтому сложно включить или отключить экспрессию гена с помощью фермента, редактирующего эпигеном, который связан с ДНК-связывающим доменом транскрипционного фактора. ДНК-связывающий домен транскрипционного фактора не подходит для нацеливания фермента, модифицирующего эпигеном. Белковый домен, способный распознать целевую последовательность ДНК размером примерно 20 п.н., сливается с каталитическим доменом из пула ферментов, модифицирующих эпигеном (так называемых эпиэффекторов), для того, чтобы получился искусственный фермент для модификации эпигенома [45].

TALE также можно ассоциировать с различными другими белками, ферментами или эпиэффекторами для повышения эффективности и точности. TALE-DNMT3ACD-3 L, TALE-CIB1; TALE/dCas9-KRAB, DNMT3ACD, DNMT3L и т. д. являются замечательными инструментами направленного метилирования с различной эффективностью. Аналогичным образом, TALE-TET1CD, TALE-TET1 и т. д. являются инструментами деметилирования, используемыми в различных исследованиях [48].

Еще одна перспективная область применения TALEN – это модификация оснований. В отличие от платформ редактирования оснований Cas9 и Cpf1, которые работают преимущественно с одноцепочечной ДНК, недавно открытая бактериальная деаминаза (DddA) катализирует дезаминирование цитидина в молекулах двухцепочечной ДНК и позволяет разрабатывать дизайнерские редакторы баз TALE-BE с альтернативными платформами нацеливания на ДНК [49]. Эти редакторы оснований использовались для создание мутаций в ядерной ДНК [50], и, в отличие от редакторов оснований на платформах Cas9 и Cpf1, в митохондриальной и хлоропластной ДНК [51], генерируя наследуемые модификации. Эти работы сделали TALE-BE первым инструментом редактирования баз, доступным для этих клеточных компартментов. TALE-BE, как и другие редакторы, основанные на TALE, должны предоставить расширенные возможности доступа к труднодоступным для редактирования локусам [52]). А используя различные правила для нацеливания на геном и взаимодействия с ним, эти редакторы откроют дополнительные сайты, выходящие за рамки ранее описанных базовых платформ редактирования Cas9 и Cpf1. [53].

СПИСОК ЛИТЕРАТУРЫ

1. Bibikova M., Golic M., Golic K.G., Carroll D. Targeted Chromosomal Cleavage and Mutagenesis in Drosophila Using Zinc-Finger Nucleases, *Genetics,* 2002. V. 161. No. 3. P. 1169–1175. doi: 10.1093/genetics/161.3.1169

2. Qasim W., Zhan H., Samarasinghe S., Adams S., Amrolia P., Stafford S., Butler K., Rivat C., Wright G., Somana K. et al. Molecular remission of infant B-ALL after infusion of universal TALEN gene-edited CAR T cells. *Science Translational Medicine*. 2017. V. 9. Article No. eaaj2013. doi: 10.1126/scitranslmed.aaj2013
3. Menz J., Modrzejewski D., Hartung F., Wilhelm R., Sprink T. Genome edited crops touch the market: a view on the global development and regulatory environment. *Front. Plant Sci*. 2020. V. 11. Article No. 586027. doi: 10.3389/fpls.2020.586027
4. Pickar-Oliver A., Gersbach C.A. The next generation of CRISPR–Cas technologies and applications. *Nat. Rev. Mol. Cell Biol.* 2019. V. 20. P. 490–507. doi: 10.1038/s41580-019-0131-5
5. Zhang B. CRISPR/Cas gene therapy. *J. Cell Physiol*. 2021. V. 236. P. 2459–2481. doi: 10.1002/jcp.30064
6. Saifaldeen M., Al-Ansari D.E., Ramotar D., Aouida M. CRISPR FokI Dead Cas9 System: Principles and Applications in Genome Engineering. *Cells*. 2020. V. 9. No. 11. Article No. 2518. doi: 10.3390/cells9112518
7. Gao H., Wu X., Chai J., Han Z. Crystal structure of a TALE protein reveals an extended N-terminal DNA binding region. *Cell Res.* 2012. V. 22. P. 1716–1720. doi: 10.1038/cr.2012.156
8. Yuan M., Ke Y., Huang R., Ma L., Yang Z., Chu Z., Xiao J., Li X., Wang S. A host basal transcription factor is a key component for infection of rice by TALE-carrying bacteria. *eLife*. 2016. V. 5. Article No. e19605. doi: 10.7554/eLife.19605
9. Moscou M.J., Bogdanove A.J. A simple cipher governs DNA recognition by TAL effectors. *Science*. 2009. V. 326. P. 1501. doi: 10.1126/science.1178817
10. Yang J., Zhang Y., Yuan P., Zhou Y., Cai C., Ren Q., Wen D., Chu C., Qi H., Wei W. Complete decoding of TAL effectors for DNA recognition. *Cell Res.* 2014. V. 24. P. 628–631. doi: 10.1038/cr.2014.19
11. Miller J., Zhang L., Xia D.F., Campo J.J., Ankoudinova I.V., Guschin D.Y., Babiarz J.E., Meng X., Hinkley S.J., Lam S.C. Improved specificity of TALE-based genome editing using an expanded RVD repertoire. *Nat. Methods*. 2015. V. 12. P. 465–471. doi: 10.1038/nmeth.3330
12. Mak A.N.S., Bradley P., Cernadas R.A., Bogdanove A.J., Stoddard B.L. The crystal structure of TAL effector PthXo1 bound to its DNA target. *Science*. 2012. V. 335. P. 716–719. doi: 10.1126/science.1216211
13. Deng D., Yan C., Pan X., Mahfouz M., Wang J., Zhu J.-K., Shi Y., Yan N. Structural basis for sequence-specific recognition of DNA by TAL effectors. *Science*. 2012. V. 335. P. 720–723 doi: 10.1126/science.1215670
14. Streubel J., Blücher C., Landgraf A. Boch J. TAL effector RVD specificities and efficiencies. *Nat. Biotechnol*. 2012. V. 30. P. 593–595 doi: 10.1038/nbt.2304
15. Becker S., Boch J. TALE and TALEN genome editing technologies. *Gene and Genome Editing*. 2021. V. 2. Article No. 100007. doi: 10.1016/j.ggedit.2021.100007
16. Boch J., Scholze H., Schornack S., Landgraf A., Hahn S., Kay S., Lahaye T., Nickstadt A., Bonas U. Breaking the Code of DNA Binding Specificity of TAL-Type III Effectors. *Science*. 2009. V. 326. No. 5959. P. 1509–1512. doi: 10.1126/science.1178811
17. Hockemeyer D., Wang H., Kiani S., Lai C.S., Gao Q., Cassady J.P., Cost G.J., Zhang L., Santiago Y., Miller J.C., et al. Genetic engineering of human pluripotent cells using TALE nucleases. *Nat. Biotechnol*. 2011. V. 29. P. 731–734. doi: 10.1038/nbt.1927
18. Guilinger J.P., Pattanayak V., Reyon D., Tsai S.Q., Sander J.D., Joung J.K., Liu D.R. Broad specificity profiling of TALENs results in engineered nucleases with improved

DNA-cleavage specificity. *Nat. Methods*. 2014. V. 11. No. 4. P. 429–435. doi: 10.1038/nmeth.2845

19. Doyle E.L., Booher N.J., Standage D.S., Voytas D.F., Brendel V.P., VanDyk J.K., Bogdanove A.J. TAL Effector-Nucleotide Targeter (TALE-NT) 2.0: tools for TAL effector design and target prediction. *Nucleic Acids Res*. 2012. V. 40. P. W117–W122. doi: 10.1093/nar/gks608
20. Grau J., Boch J., Posch S. TALENoffer: genome-wide TALEN off-target prediction. *Bioinformatics*. 2013. V. 29. P. 2931–2932. doi: 10.1093/bioinformatics/btt501
21. Cong L., Zhou R., Kuo Y.C., Cunniff M., Zhang F. Comprehensive interrogation of natural TALE DNA-binding modules and transcriptional repressor domains. *Nat. Commun*. 2012. V. 3. Article No. 968. doi: 10.1038/ncomms1962
22. Richter A., Streubel J., Blücher C., Szurek B., Reschke M., Grau J., Boch J. A TAL effector repeat architecture for frameshift binding. *Nat. Commun*. 2014. V. 5. Article No. 3447. doi: 10.1038/ncomms4447
23. Sakuma T., Ochiai H., Kaneko T., Mashimo T., Tokumasu D., Sakane Y., Suzuki K., Miyamoto T., Sakamoto N., Matsuura S., Yamamoto T. Repeating pattern of non-RVD variations in DNA-binding modules enhances TALEN activity. *Sci. Rep*. 2013. V. 3. Article No.3379. doi: 10.1038/srep03379
24. Sakuma T., Yamamoto T. Engineering Customized TALENs Using the Platinum Gate TALEN Kit. *Methods Mol. Biol*. 2016. V.1338. P. 61–70. doi: 10.1007/978-1-4939-2932-0_6
25. Xue J., Lu Z., Liu W., Wang S., Lu D., Wang X., He X. The genetic arms race between plant and *Xanthomonas*: lessons learned from TALE biology. *Sci. China Life Sci*. 2021. V. 64. No. 1. P. 51–65. doi: 10.1007/s11427-020-1699-4
26. Streubel J., Blücher C., Landgraf A., Boch J. TAL effector RVD specificities and efficiencies. *Nat. Biotechnol*. 2012. V. 30. P. 593–595. doi: 10.1038/nbt.2304
27. Čermák T., Doyle E.L., Christian M., Wang L., Zhang Y., Schmidt C., Baller J.A., Somia N.V., Bogdanove A.J., Voytas D.F. Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic Acids Res*. 2011. V. 39. Article No. e82. doi: 10.1093/nar/gkr218
28. Balwierz P.J., Carninci P., Daub C.O., Kawai J., Hayashizaki Y., Van Belle W., Beisel C., van Nimwegen E. Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. *Genome Biol*. 2009. V. 10. Article No. R79. doi: 10.1186/gb-2009-10-7-r79
29. Bradley D., Roth G. Adaptive Thresholding using the Integral Image. *Journal of Graphics GPU and Game Tools*. 2007. V. 12. P. 13–21. doi: 10.1080/2151237X.2007.10129236
30. Umer M., Herceg Z. Deciphering the epigenetic code: an overview of DNA methylation analysis methods. *Antioxid Redox Signal*. 2013. V. 18. P. 1972–1986. doi: 10.1089/ars.2012.4923
31. Jabbari K., Bernardi G. Cytosine methylation and CpG, TpG (CpA) and TpA frequencies. *Gene*. 2004. V. 333. P. 143–149. doi: 10.1016/j.gene.2004.02.043
32. Herman J.G., Graff J.R., Myöhänen S., Nelkin B.D., Baylin S.B. Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. *Proc. Natl. Acad. Sci. U.S.A*. 1996. V. 93. No. 18. P. 9821–9826. doi: 10.1073/pnas.93.18.9821
33. Тетуев Р.К., Ольшевец М.М., Erman B., Atilgan C. Модель "сломанная молния": предсказание сайтов связывания TALEN'ов на основе скользящего среднего. В: Доклады международной конференции «Математическая биология и биоинформатика». Ред.: Лахно В.Д. Т. 6. Пущино, 2016. С. 70–71.

34. Li L., Piatek M.J., Atef A., Piatek A., Wibowo A., Fang X., Sabir J.S.M., Zhu J.-K., Mahfouz M.M. Rapid and highly efficient construction of TALE-based transcriptional regulators and nucleases for genome modification. *Plant Mol. Biol*. 2012. V. 78. P. 407–416. doi: 10.1007/s11103-012-9875-4
35. Fine E.J., Cradick T.J., Zhao C.L., Lin Y., Bao G. An online bioinformatics tool predicts zinc finger and TALE nuclease off-target cleavage. *Nucleic Acids Res*. 2013. V. 42. Article No. e42. doi: 10.1093/nar/gkt1326
36. Heigwer F., Kerr G., Walther N., Glaeser K., Pelz O., Breinig M., Boutros M. E-TALEN: a web tool to design TALENs for genome engineering. *Nucleic Acids Res*. 2013. V. 41. Article No. e190. doi: 10.1093/nar/gkt789
37. Neff K.L., Argue D.P., Ma A.C., Lee H.B., Clark K.J., Ekker S.C. Mojo Hand, a TALEN design tool for genome editing applications. *BMC Bioinform*. 2013. V. 14. Article No. 1. doi: 10.1186/1471-2105-14-1
38. Montague T.G., Cruz J.M., Gagnon J.A., Church G.M., Valen E. CHOPCHOP: a CRISPR/Cas9 and TALEN web tool for genome editing. Nucleic Acids Res. 2014. V. 42. P. W401–W407. doi: 10.1093/nar/gku410
39. Jensen T.L., Gøtzsche C.R., Woldbye D.P.D. Current and Future Prospects for Gene Therapy for Rare Genetic Diseases Affecting the Brain and Spinal Cord. *Front. Mol. Neurosci*. 2021. V. 14. Article No. 695937. doi: 10.3389/fnmol.2021.695937
40. Kaiser J. A gentler way to tweak genes: Epigenome editing. *Science*. 2022. V. 376. P. 1034–1035. doi: 10.1126/science.add2703
41. Margueron R., Reinberg D. Chromatin structure and the inheritance of epigenetic information. *Nat. Rev. Genet.* 2010. V. 11. P. 285–296. doi: 10.1038/nrg2752
42. Allis C.D., Jenuwein T. The molecular hallmarks of epigenetic control. *Nat. Rev. Genet*. 2016. V. 17. P. 487–500. doi: 10.1038/nrg.2016.59
43. Назипова Н.Н. Разнообразие некодирующих РНК в геномах эукариот. *Мат. биология и биоинформатика*. 2021. Т. 16. N 2. С. 256–298. doi: 10.17537/2021.16.256
44. Cook P.R. A model for all genomes: The role of transcription factories. *J. Mol. Biol*. 2010. V. 395. P. 1–10. doi: 10.1016/j.jmb.2009.10.031
45. Ueda J., Yamazaki T., Funakoshi H. Toward the Development of Epigenome Editing-Based Therapeutics: Potentials and Challenges. *International Journal of Molecular Sciences*. 2023. V. 24. No.5. Article No. 4778. doi: 10.3390/ijms24054778
46. Baker M.P., Reynolds H.M., Lumicisi B., Bryson C.J. Immunogenicity of protein therapeutics: The key causes, consequences and challenges. *Self/Nonself*. 2010. V. 1. P. 314–322. doi: 10.4161/self.1.4.13904
47. de Groote M.L., Verschure P.J., Rots M.G. Epigenetic Editing: Targeted rewriting of epigenetic marks to modulate expression of selected target genes. *Nucleic Acids Res*. 2012. V. 40. P. 10596–10613. doi: 10.1093/nar/gks863
48. Lei Y., Huang Y.H., Goodell M.A. DNA methylation and de-methylation using hybrid site-targeting proteins. *Genome Biol*. 2019. V. 19. Article No. 187. doi: 10.1186/s13059-018-1566-2
49. Mok B.Y., de Moraes M.H., Zeng J., Bosch D.E., Kotrys A.V., Raguram A., Hsu F., Radey M.C., Peterson S.B., Mootha V.K. et al. A bacterial cytidine deaminase toxin enables CRISPR-free mitochondrial base editing. *Nature*. 2020. V. 583. P. 631–637. doi: 10.1038/s41586-020-2477-4
50. Mok B.Y., Kotrys A.V., Raguram A., Huang T.P., Mootha V.K., Liu D.R. CRISPR-free base editors with enhanced activity and expanded targeting scope in mitochondrial and nuclear DNA. *Nat. Biotechnol*. 2022. V. 40. P. 1378–1387. doi: 10.1038/s41587-022-01256-8

51. Kang B.C., Bae S.J., Lee S., Lee J.S., Kim A., Lee H., Baek G., Seo H., Kim J., Kim J.-S. Chloroplast and mitochondrial DNA editing in plants. *Nat. Plants*. 2021. V. 7. P. 899–905. doi: 10.1038/s41477-021-00943-9

52. Jain S., Shukla S., Yang C., Zhang M., Fatma Z., Lingamaneni M., Abesteh S., Lane S.T., Xiong X., Wang Y., et al. TALEN outperforms Cas9 in editing heterochromatin target sites. *Nat. Commun*. 2021. V. 12. P. 606–610. doi: 10.1038/s41467-020-20672-5

53. Boyne A., Yang M., Pulicani S., Feola M., Tkach D., Hong R., Duclert A., Duchateau P., Juillerat A. Efficient multitool/multiplex gene engineering with TALE-BE. *Front. Bioeng. Biotechnol*. 2022. V. 10. Article No. 1033669. doi: 10.3389/fbioe.2022.1033669

Рукопись поступила в редакцию 13.11.2023, переработанный вариант поступил 19.11.2023.
Дата опубликования 31.12.2023.

========== **BIOINFORMATICS** ==========

Statistical Model for Predicting TALEN-DNA Binding Sites Based On Moving Average

Tetuev R.K., Nazipova N.N.

Institute of Mathematical Problems of Biology RAS, Keldysh Institute of Applied Mathematics RAS, Pushchino, Russia

Abstract. In this paper, we propose a new approach to the in-silico prediction of any possible DNA binding sites for the user-defined artificial TALENs. This approach based on the exponential moving average model and developed as an online service TANDIS. The direct validation of our prediction model based on the direct matching with the known results of the certain in-vitro experiments, while for the verification of its accuracy we use comparative analysis against other similar popular services like TALE-NT and TALENoffer. So thus, we have found out that the exponential moving average model brings very good results comparable with those of the Markov chain model used in TALENoffer, but TANDIS can do it much more easily because its model is much simpler. The TALE-NT prediction is even faster than ours for it has an utmost simple position-independent scoring system and drastically simplified filtering rules for the case of paired TALEs, which makes however, on the other hand, the results of such TALE-NT 's prediction much less competitive. Besides being the compromise between accuracy and efficiency, the exponential moving average model has only five parameters, so in future, it could be easily used for more intense prediction, and probably later, it can be used to cast some light on our understanding of real physical principles of the attractive interaction between a certain TALE and a random DNA site.

***Key words**: genome editing, TALEN, exponentially weighted moving average, in-silico binding site prediction, online server.*