Информационные процессы, Том 12, № 3, 2012, стр. 229–247.

Моделирование первичной специфичности протеаз семейства MMP методами машинного обучения

Г.Г. Федонин, М.Д. Казанов

Федеральное государственное бюджетное учреждение науки
Институт проблем передачи информации им. А.А.Харкевича
Российской академии наук, Москва, Россия

Поступила в редколлегию 12.07.2012

Аннотация—Исследовались восемь протеаз из разных подгрупп семейства MMP. На основе измерений протеолитической активности для 1369 пептидов, отобранных экспериментально из 64 миллионов пептидов длиною шесть аминокислот с использованием фагового дисплея, и хорошо разрезаемых хотя бы одной из протеаз, были построены предсказательные модели первичной специфичности этих протеаз.

Для предсказания протеолитической активности каждой протеазы на пептиде по его аминокислотной последовательности использовались линейные регрессионные модели и метод ближайших соседей (kNN). Наивный байесовский классификатор, логистическая регрессия и метод ближайших соседей использовались для решения бинарной задачи классификации, в которой пептиды с известной активностью считались 'положительными', а случайные пептиды — 'отрицательными'.

Для определения релевантных для предсказания специфичности позиций аминокислотной последовательности пептида использовался жадный отбор признаков.

В работе предложен метод сокращения размерности пространства признаков, основанный на параметризации аминокислот, и показано, что его использование повышает качество предсказания регрессионных и классификационных моделей.

Предложен специальный алгоритм обучения линейной регрессионной модели: к обучающим пептидам добавляются случайные пептиды. Функционал качества для случайных пептидов полагался равным нулю, если предсказанное значение было меньше минимального значения активности в выборке для данной протеазы, и, в противном случае, равным квадрату отклонения предсказанного значения от минимальной активности (также как и в стандартном алгоритме наименьших квадратов OLS).

Модели были проверены на CutDB — базе экспериментально зафиксированных событий протеолиза. Для демонстрации эффективности предложенных моделей были построены ROC-кривые с использованием данных из CutDB.

КЛЮЧЕВЫЕ СЛОВА: биоинформатика, протеазы, матриксные металлопротеиназы, первичная специфичность протеаз, машинное обучение, сокращение размерности пространства признаков.

1. ВВЕДЕНИЕ

Протеазы (или протеиназы) — ферменты, катализирующие реакцию расщепления пептидной связи (протеолиза). Матриксные металлопротеиназы (MMP) — семейство цинк-зависимых эндопептидаз, проявляющих свою каталитическую активность по отношению к мембранным белкам, белкам секреторных путей и белкам межклеточного пространства.

Основной и единственной функцией MMP долгое время считалось деградация межклеточного матрикса [1, 2]. Новое видение биологических функций MMP включает их участие

в таких процессах как репарация тканей, ангиогенез, процессах иммунного ответа, развитии опухолей и воспалительных процессах [3, 4]. Установление новых функций ММР стало возможным путем определения новых субстратов ММР — цитокинов, хемокинов, рецепторов и антибактериальных пептидов [5].

Текущий каталог субстратов ММР несомненно является далеко не полным, а трудоемкость применения экспериментальных техник делает разработку биоинформатических методов предсказания субстратов актуальной задачей, имеющей непосредственную практическую важность в области медицины и разработки лекарств [6, 7, 8].

Рентгеновская кристаллография структур каталитических доменов нескольких матриксных металлопротеиназ [9, 10, 11, 12, 13, 14] показала, что этот домен имеет форму сплюснутой сферы размера $35 \times 30 \times 30$ Å ($3.5 \times 3 \times 3$ нм). Активный сайт каталитического домена представляет собой канавку размером 20 Å (2 нм), простирающуюся поперек каталитического домена. В области каталитического домена, образующей активный сайт, находится важный для катализа ион Zn^{2+}, который связан с тремя остатками гистидина. Указанные остатки содержатся в консервативной последовательности HExxHxxGxxH, являющейся, следовательно, цинк-связывающим мотивом [15]. Мутация любого из этих остатков гистидина приводит к исчезновению каталитической активности [16, 17].

Связывание главной цепи субстрата с активным сайтом является важным элементом процесса протеолитического катализа. Данное связывание происходит за счет формирования структуры антипараллельного β-листа с формированием водородных связей между субстратом и ферментом. Так называемый карман специфичности, расположенный справа и слева от активного сайта, связывается с боковыми цепями аминокислотных остатков субстрата, тем самым определяя первичную специфичность протеазы [18]. Аминокислотные остатки кармана, расположенные по обе стороны от активного центра называются субсайтами. Остатки расположенные слева от активного сайта (см. рис. 1) обозначаются как S_i, где i — номер остатка, считая от активного центра. Остатки расположенные справа от активного сайта обозначаются как S'_i, где i — номер остатка, считая от активного центра. Позиции сайта связывания субстрата, соответствующие позициям в кармане, обозначаются как P_i и P'_i соответственно.

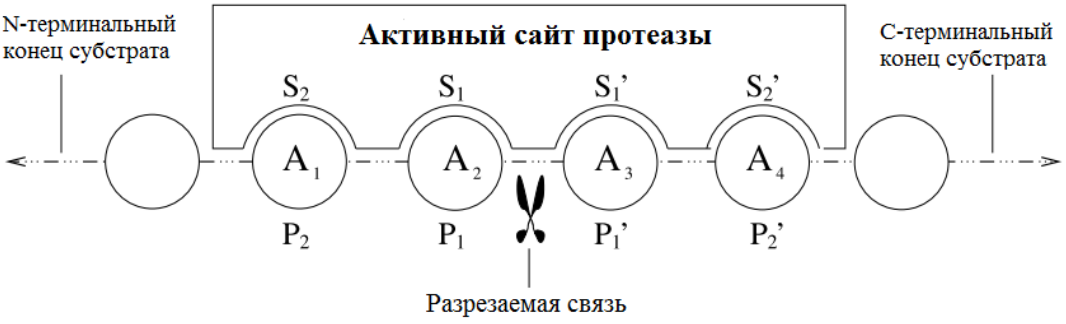

Рис. 1. Диаграмма взаимодействия активного сайта гипотетической протеазы, содержащего четыре субсайта и аминокислот $(A_1, \ldots, A_4)$ субстрата. Показана нотация Schechter и Berger [19], для субсайтов $(S_2, \ldots, S'_2)$ и остатков $(P_2, \ldots, P'_2)$ относительно разрезаемой связи.

Последовательности известных субстратов различных протеаз, а также результаты экспериментов по описанию первичной специфичности использовались разными авторами для построения предсказательных моделей первичной специфичности протеаз — консенсусных мотивов или позиционных весовых матриц (PSSM) [20, 21, 22, 23, 24, 25, 26, 27]. Наряду с частотными матрицами, многими авторами применялись различные методы машинного обучения [28, 29].

Передовой разработкой в области статистического предсказания субстратов на основе первичной специфичности был PeptideCutter — сервер, предсказывающий сайты разрезания и субстраты протеаз различных семейств. Из-за недостатка данных PeptideCutter основывался на данных о специфичности определенных каспаз [20]. Lohmuller и соавт. [21] разработали инструмент для предсказания субстратов пептидаз (PEPS), основанный на позиционных весовых матрицах (PSSM) сайтов разрезания субстратов катепсина B, катепсина L и каспазы-3. Garay-Malpartida и соавт. [25] разработали программу CasPredictor, показавшую улучшение по сравнению с ранее использованными методами, достигнув точности в 81% на наборе из 137 экспериментально подтвержденных сайтах разрезания. Эта программа анализирует частоты аминокислот и наличие в окрестности разреза «PEST»-последовательностей [30,31]. Программа GraBCas Backes и соавт. [26] улучшила предыдущие модели на основе PSSM за счет использования данных, полученных с помощью библиотек пептидов. PoPS, предложенный Boyd и соавт. в [27], дополняет модель PSSM вектором весов, характеризующим вклад субсайтов в специфичность, и корректирующими правилами, позволяющими учесть кооперативность.

Для предсказания сайтов разрезания каспаз Yang [28] экспериментировал с различными нейронными сетями (однослойным персептроном, многослойным персептроном), машиной опорных векторов (SVM) и предложенной автором сетью базисных функций, основанных на попарной близости, с выбором весов функций методом Байеса с использованием равномерного и нормального распределений, а также смеси нормальных распределений. Автор достиг точности 97% на маленькой выборке из 13 последовательностей. Машины опорных векторов применялись Wee и соавт. [29] также для предсказания сайтов каспаз.

В настоящей работе исследовались восемь протеаз из разных подгрупп семейства ММР, для которых были доступны количественные экспериментальные данные об эффективности реакции протеолиза, предоставленные лабораторией Д. Смита Медицинского исследовательского института Сэнфорда-Бернема. Данные использовались для построения моделей первичной специфичности исследованных ММР методами машинного обучения.

2. МАТЕРИАЛЫ И МЕТОДЫ

2.1. Подготовка данных

В настоящей работе использовались данные эксперимента по профилированию специфичности протеаз методом фагового дисплея [32], предоставленные группой Д. Смита Медицинского исследовательского института Сэнфорда-Бернема. Экспериментаторы выбрали 8 членов семейства, относящихся к трем филогенетически удаленным друг от друга подгруппам:

1. гликозилфосфатидилинозитол(GPI)-заякоренные (ММР-17 и ММР-25)
2. желатиназы (ММР-2 и ММР-9)
3. трансмембранные (ММР-14, ММР-15, ММР-16 и ММР-24).

Субстратный фаговый дисплей проводился путем отображения случайных пептидов в составе гибридного белка с белком g3p (gene 3 protein) нитеобразного бактериофага M13. Использовалась библиотека пептидов, содержащая всевозможные гексапептиды. Из полученных в эксперименте пептидов были отобраны 1369, каждый из которых хорошо разрезался хотя бы одной из протеаз. Для этих пептидов была измерена эффективность разрезания — усиление гидролиза в образцах, обработанных протеазой. Усиление гидролиза (или Hydrolysis Score, HS) определялось как единица минус отношение OD образцов, обработанных протеазой, и соответствующих необработанных образцов. Так как концентрация субстратов (10^{-10} M) была сильно ниже типичной KM (порядка 10^{-4} M для пептидных субстратов), значения $kcat/KM$ были определены с помощью интегрального уравнения Михаэлиса-Ментен:

$$[P] = [S]_0(1 - \exp^{-kt}),$$

где $[P]$ — концентрация продукта, $[S]_0$ — начальная концентрация субстрата, $k = \frac{kcat/KM}{[E]}$, $[E]$ — концентрация фермента.

$$[P] = [S]_0 - [S]_t,$$

где $[S]_t$ — концентрация субстрата в момент t.

$$1 - \frac{[S]_t}{[S]_0} = HS = 1 - \exp^{-kt} \Rightarrow kt = -\ln(1 - HS)$$

$$k = \frac{kcat/KM}{[E]} \Rightarrow \frac{kcat}{KM} = -\frac{\ln(1 - HS)}{t[E]}$$

После идентификации субстратов были прочитаны последовательности гексапептидов. Для секвенирования разрезанных пептидов использовалась ПЦР-амплификация участка g3p, содержащего случайную вставку, с последующим секвенированием ампликонов. Для установления места разреза использовался времяпролетный MALDI-масс-спектрометр (TOF MS).

Отобранные пептиды были выравнены по положению разреза и дополнены, при необходимости, фаговыми аминокислотами, которые являлись смежными с данным пептидом в процессе фагового дисплея. Таким образом, было получено выравнивание аминокислотных последовательностей пептидов длиной 10 а.о., с сопоставленными каждой последовательности значениями эффективности протеолиза.

В настоящей работе для задач классификации и ранжирования также использовалась дополнительная выборка случайных пептидов. Сначала были порождены случайные пептиды длиной 6 а.о. с равномерным на множестве аминокислот распределением в каждой позиции пептида. Позиции пептида при этом считались независимыми. Пептиды, совпадающие с пептидами из выравнивания, были исключены. Случайные пептиды длины 10 получали в два этапа: сначала генерировались случайные пептиды длины 6, а потом каждый из полученных пептидов длины 6 дополнялся по краям соответствующими фаговыми аминокислотами всеми возможными способами. Эта процедура необходима для устранения краевых эффектов в выборке пептидов длины 10: при фаговом дисплее воздействию протеаз подвергались всевозможные пептиды длины 6, и потому, не все возможные последовательности длины 10 были протестированы. Следовательно, в полученном выравнивании экспериментальных пептидов частоты фаговых аминокислот в краевых позициях завышены.

Для независимой проверки качества построенных в работе моделей первичной специфичности протеаз использовалась база данных CutDB ($http://cutdb.burnham.org$) [33]. Эта база — одна из первых попыток систематизировать документированные протеолитические события для природных субстратов *in vivo* или *in vitro*. На данный момент эта база данных содержит более 11 000 событий для более 600 протеаз в более чем 3000 субстратов. Вся информация получена из публичных архивов (таких как MEROPS и HPRD) и на основе анализа публикаций. Из исследованных восьми протеаз только для трех в CutDB имеется достаточное число событий протеолиза: MMP9 (334 события), MMP2 (135 событий) и MMP14 (89 событий). Эти данные использовались для построения ROC-кривых для оценки качества и сравнения разработанных моделей.

2.2. Постановки задач и оценка качества алгоритмов

Задача моделирования первичной специфичности может быть поставлена по-разному. В настоящей работе были построены модели для трех различных постановок задачи. Первая — задача регрессии: для каждой протеазы построить модель, предсказывающую эффективность разрезания по аминокислотной последовательности пептида. Такая постановка позволяет использовать информацию о числовых значениях эффективности небольшого числа пептидов,

для которых эти значения известны, однако при этом теряется информация о том, что почти все остальные пептиды разрезаются с меньшей эффективностью или не разрезаются вовсе.

Вторая постановка — задача классификации пептидов: для каждой протеазы построить модель, классифицирующую пептиды по аминокислотной последовательности на эффективно и неэффективно разрезаемые. Эти модели обучаются одновременно на пептидах с известными значениями эффективности (положительный класс) и случайных пептидах (отрицательный класс), а числовые значения эффективности не используются.

Третья постановка — задача ранжирования пептидов: для каждой протеазы построить модель, позволяющую упорядочить пару пептидов по эффективности, имея их аминокислотные последовательности. Эти модели обучаются одновременно на пептидах с известными значениями эффективности (положительный класс) и случайных пептидах (отрицательный класс), и числовые значения эффективности тоже используются: можно упорядочить пары 'положительный–положительный' и 'положительный–отрицательный', тогда как порядок пар 'отрицательный–отрицательный' неизвестно. Ошибка ранжирования может быть вычислена и для регрессионных, и для классификационных моделей.

Для оценки качества алгоритмов использовалась 8×10–кросс-валидация(8×10–fold cross-validation). Выборка случайным образом разбивалась на 8 частей и каждая часть поочередно считалась тестовой выборкой, а объединение остальных — обучающей. Модели оптимизировались на обучающей выборке, а на тестовой — вычислялась соответствующая мера качества модели. Значения, полученные для разных разбиений, усреднялись. Весь процесс, включая разбиение на блоки, повторялся десять раз и результаты усреднялись по всем итерациям.

Качество регрессионных моделей оценивалось среднеквадратичной ошибкой прогноза:

$$Err = \sqrt{\sum_{i=1}^{n} (y_i - f(x_i))^2 / n} \ ,$$

где n — число аминокислотных последовательностей пептидов, y_i — значение эффективности разрезания i-го пептида, x_i — аминокислотная последовательность i-го пептида, $f(x_i)$ — прогноз модели для i-го пептида.

Качество моделей-классификаторов оценивалось средней частотой ошибочной классификации объектов обоих классов. Количество пептидов с известными значениями эффективности фиксировано, а количество случайных пептидов можно менять. Чтобы оценка качества моделей классификации не зависела от количества случайных пептидов, ошибка классификации считалась по каждому классу отдельно, а результаты усреднялись:

$$Err = \frac{\sum_{i=1}^{n_+} [f(x_i)]_-}{2n_+} + \frac{\sum_{i=1}^{n_-} [f(x_i)]_+}{2n_-} \ ,$$

где n_- — число случайных пептидов, n_+ — число хорошо разрезаемых пептидов, x_i — аминокислотная последовательность i-го пептида, $f(x_i) \in \{-1, 1\}$ — прогноз модели для i-го пептида, $[a]_- = 1$ при $a < 0$, $[a]_- = 0$ иначе, $[a]_+ = 1$ при $a > 0$, $[a]_+ = 0$ иначе.

Качество ранжирующих моделей оценивалось средней частотой ошибок, причем оценка считалась отдельно по всем парам положительных пептидов и по всем парам положительных пептидов со всеми отрицательными, а результаты усреднялись. При этом считалось, что любой положительный пептид разрезается более эффективно, чем любой отрицательный:

$$Err = \frac{\sum_{i=1}^{n_+} \sum_{j=1}^{n_+} [(y_i - y_j)(f(x_i) - f(x_j))]_-}{2n_+(n_+ - 1)} + \frac{\sum_{i=1}^{n_+} \sum_{j=1}^{n_-} [f(x_i) - f(x_j)]_-}{2n_+ n_-} \ ,$$

где n_- — число случайных пептидов, n_+ — число хорошо разрезаемых пептидов, x_i — аминокислотная последовательность i-го пептида, $f(x_i)$ — прогноз модели для i-го пептида, $[a]_- = 1$ при $a < 0$, $[a]_- = 0$ иначе.

Помимо кросс-валидации, алгоритмы тестировались на независимой выборке, полученной из CutDB. Для этого из всех последовательностей белковых субстратов, для которых в CutDB зафиксировано хотя бы одно событие протеолиза хотя бы для одной из исследованных протеаз, были получены все содержащиеся в них гексапептиды. Пептиды, содержащие разрез в центре, были помечены как положительные, остальные — как отрицательные.

Все тестируемые алгоритмы обучались на полной обучающей выборке — выравнивании пептидов (в т.ч. случайных) с приписанными им значениями эффективности (для регрессионных алгоритмов) или метками классов (для классификаторов), из которого были удалены все пептиды, встречающиеся в выборке из CutDB. Обученные алгоритмы использовались для сортировки пептидов из CutDB по убыванию предсказанной эффективности разрезания. Для каждого алгоритма были построены ROC-кривые и вычислялась площадь под ROC кривой (AUC), характеризующая общее качество алгоритма.

2.3. Алгоритмы

При обучении классификационных моделей пептиды обучающей выборки взвешивались обратно пропорционально числу пептидов данного класса.

Наивный байесовский классификатор

Байесовский классификатор [34] оценивает вероятности наблюдать у данной последовательности класс c по формуле Байеса:

$$P(c_i|S) = \frac{P(c_i)P(x|c_i)}{\sum_j P(c_j)P(x|c_j)} \ ,$$

где c_i — i-й класс (тип нуклеотида n или класс пептида), x — аминокислотная последовательность пептида, $P(c)$ — априорная вероятность появления класса c. Априорные вероятности классов считались равными.

Наивный байесовский классификатор предполагает, что позиции в А.П. условно независимы в совокупности, т.е.

$$P(x|c) = \prod_i P(a_i|c) \ ,$$

где a_i — аминокислотный остаток в позиции i. Вероятности $P(a_i|c)$ оценивали по соответствующим частотам выборки.

В приложениях задачи бинарной классификации часто возникают, когда нужно отделить небольшую выборку положительных объектов от всевозможных остальных. Задача предсказания сайтов разрезания протеаз — одна из таких задач. Наивный байесовский классификатор в данном случае эквивалентен позиционным весовым матрицам (PWM, PSSM): эти матрица могут быть получены логарифмированием функции правдоподобия положительного класса и наоборот. Функции правдоподобия отрицательного класса могут быть оценены с высокой точностью с помощью равномерного распределения, и являются постоянными для всех пептидов.

Линейная регрессия

Для предсказания численных значений эффективности разрезания пептидов использовалась классическая многомерная линейная регрессия [35]. Предполагается, что обучающие объекты задаются k-мерными векторами числовых признаков $x = (x_1, \ldots, x_k)$. Эффективность

предполагается линейно зависящей от признаков: алгоритм строит линейную прогнозирующую функцию. Каждому числовому признаку присваивается коэффициент. При прогнозировании вычисляется линейная комбинация признаков объекта.

При обучении алгоритм минимизирует квадратичную функцию потерь на обучающей выборке. Для повышения устойчивости полученного решающего правила применялась регуляризация Тихонова (известная также как гребневая регрессия): к квадратичной функции потерь добавлялось слагаемое, штрафующее сумму квадратов коэффициентов обучаемой линейной функции:

$$L(\vec{\alpha}) = \sum_{i=1}^{n} (y_i - \langle \vec{\alpha}, \vec{x}_i \rangle)^2 + \beta \sum_{i=1}^{k} \alpha_i^2 \to \min_{\vec{\alpha}} \ ,$$

где n — число объектов обучающей выборки, $\vec{\alpha}$ — вектор коэффициентов, x_i — вектор признаков i-го объекта, $\beta \sum_{i=1}^{k} \alpha_i^2$ — регуляризующее слагаемое, β — априори зафиксированный параметр регуляризации.

Линейная регрессия предполагает, что объекты характеризуются числовыми признаками. В нашем случае признаки номинальные. Использовали два подхода. Первый — стандартный метод бинаризации таких признаков: каждому i-му аминокислотному остатку был сопоставлен свой признак-индикатор: $f_i(a) = 1$, когда $a = a_i$, и $f_i(a) = 0$ в остальных случаях. Некоторые аминокислоты в некоторых позициях пептидного выравнивания встречаются слишком редко. Коэффициент, приписываемый соответствующему бинарному признаку, будет неустойчив при вариации выборки.

Качество прогнозов алгоритма можно повысить, объединив в каждой позиции пептида аминокислотные остатки, содержащиеся в менее чем в заранее определенном числе пептидов, в одну группу, т.е. приписать всем этим аминокислотам один и тот же признак-индикатор. Пороговое число пептидов определяется экспериментально. Второй подход — параметризация аминокислотных остатков — будет описан ниже.

Логистическая регрессия

Логистическая регрессия [36] — линейный бинарный классификатор. Предполагается, что обучающие объекты задаются k-мерными векторами числовых признаков $x = (x_1, \ldots, x_k)$ с бинарной меткой класса $y \in \{-1, 1\}$. Каждому числовому признаку при обучении присваивается вес.

Решающее правило — линейное:

$$f(x_1, \ldots, x_k) = \operatorname{sign}\Big(\sum_{i=1}^{k} \alpha_i x_i\Big) \ ,$$

или в векторной форме:

$$f(\vec{x}) = \operatorname{sign}(\langle \vec{\alpha}, \vec{x} \rangle) \ ,$$

где α_i — вес i-го признака, x_i — значение i-го признака.

Обучение классификатора состоит в подборе весов путем максимизации функционала качества на обучающей выборке. Для повышения качества классификатора использовался метод регуляризации Тихонова. Функционал качества имеет вид:

$$L(\vec{\alpha}) = \sum_{i=1}^{n} w_i \ln \sigma(y_i \langle \vec{\alpha}, \vec{x}_i \rangle) - \beta \sum_{i=1}^{k} \alpha_i^2 \to \max_{\vec{\alpha}} \ ,$$

где индекс i пробегает объекты обучающей выборки, y_i — класс i-го обучающего объекта, $\sigma(z) = 1/(1+exp(-z))$ — логистическая (сигмоидная) функция, w_i — веса объектов, $\beta \sum_{i=1}^{k} \alpha_i^2$ — регуляризующее слагаемое, β — коэффициент регуляризации.

Вероятности наблюдать каждый из классов у объекта с данным вектором признаков вычисляются по формуле:

$$P(y) = \frac{1}{1 + \exp^{-y_i \langle \vec{\alpha}, \vec{x}_i \rangle}}$$

где y_i — класс, $\vec{x}_i$ — вектор признаков, $\vec{\alpha}$ — вектор весов.

Как и в случае с линейной регрессией, требуется заменить номинальные признаки числовыми: использовались те же методы.

Метод k ближайших соседей

Метод k ближайших соседей (k nearest neighbors, kNN), применяемый обычно для решения задач классификации, основан на использовании меры близости на множестве объектов. Обучение алгоритма состоит в запоминании обучающей выборки. В процессе классификации, в простейшем случае, при $k = 1$, новому объекту приписывается класс ближайшего объекта обучающей выборки. Часто можно повысить качество классификации, определяя класс нового объекта голосованием k ближайших к нему объектов обучающей выборки. Простое голосование можно заменить взвешиванием ответов объектов пропорционально их близости к классифицируемому объекту.

В задаче классификации в качестве меры близости использовалась доля совпавших аминокислотных остатков в последовательностях пептидов. Использовалось взвешенное голосование. Для увеличения вклада более близких соседей значение близости возводилось в различные целые степени. Решающее правило имело вид:

$$f(\vec{x}) = \text{sign}(\sum_{j=1}^{k} s_j^n c_j) \ ,$$

где s_j — близость j-ого ближайшего пептида к пептиду, для которого строится прогноз, $c_j \in \{-1, 1\}$ — тип j-ого ближайшего пептида (разрезается или нет). Наилучшая точность была достигнута при $k = 20, n = 2$.

Такой же алгоритм (но без функции sign() в решающем правиле) использовался в задаче ранжирования: пептиды упорядочивались в соответствии с полученным значением линейной комбинации. Максимальная точность достигалась при $k = 15, n = 3$.

Взвешенное голосование ближайших соседей можно использовать и для решения задачи регрессии. При этом в качестве прогноза используется взвешенное среднее значений прогнозируемой функции на ближайших объектах. Использовалась также мера близости, что и в задаче классификации пептидов. Для увеличения вклада более близких соседей значение близости возводилось в различные целые степени. Прогнозируемое значение рассчитывалось так:

$$y(\vec{x}) = \frac{\sum_{j=1}^{k} s_j^n y_j}{\sum_{j=1}^{k} s_j^n} \ ,$$

где s_j — близость j-ого ближайшего пептида к пептиду, для которого строится прогноз, y_j — значение активности рассматриваемой протеазы на j-ом ближайшем пептиде. Наилучшая точность была достигнута при $k = 15, n = 3$.

Параметризация аминокислот

В настоящей работе в задачах прогнозирования активности протеаз и классификации субстратных пептидов использовалась техника сокращения размерности пространства признаков, основанная на описании аминокислот числовыми признаками, значения которых не зависят о позиции аминокислоты в пептиде и оптимизируются на обучающей выборке вместе с другими параметрами модели. При этом тип оптимизируемого на обучающей выборке функционала качества (квадратичного или логистического) не меняется.

Пусть аминокислоты характеризуются k параметрами. Тогда линейная функция от пептида будет иметь вид:

$$f(x_1, \ldots, x_l) = \sum_{i=1}^{l} \sum_{j=1}^{k} \alpha_{ij} f_j(x_i) \ ,$$

где l — длина пептида, k — число параметров, α_{ij} — коэффициент при j-том параметре i-ой позиции пептида, $f_j(x_i)$ — значение j-го параметра аминокислотного остатка, стоящего в i-ой позиции пептида.

Как показали эксперименты, лучшие результаты достигаются, если параметры аминокислот одинаковы для всех протеаз и оптимизируются одновременно, т.е. оптимизируется сумма соответствующих функционалов качества всех протеаз. Лучшие результаты в задаче классификации достигались при $k = 4$, в задаче прогнозирования активности — при $k = 7$.

Оптимизируемый функционал линейной регрессии имеет вид:

$$L(\vec{\alpha}_1, \ldots, \vec{\alpha}_m) = \sum_{s=1}^{m} (\sum_{i=1}^{n_s} (y_{si} - \sum_{p=1}^{l} \sum_{j=1}^{k} \alpha_{spj} f_j(x_{ip}))^2 + \beta \sum_{p=1}^{l} \sum_{j=1}^{k} \alpha_{spj}^2) \to \min_{\vec{\alpha}_1, \ldots, \vec{\alpha}_m} \ ,$$

где m — число протеаз, n_s — количество пептидов с известными значениями активности для s-ой протеазы, y_{si} — активность s-ой протеазы на i-ом пептиде, l — длина пептида, k — число параметров, α_{spj} — коэффициент при j-ом параметре p-ой позиции пептида для s-ой протеазы, $f_j(x_{ip})$ — значение j-го параметра аминокислотного остатка, стоящего в p-ой позиции i-го пептида, β — коэффициент регуляризации.

Оптимизируемый функционал логистической регрессии имеет вид:

$$L(\vec{\alpha}_1, \ldots, \vec{\alpha}_m) = \sum_{s=1}^{m} (\sum_{i=1}^{n_s} \ln \sigma(y_{si} \sum_{p=1}^{l} \sum_{j=1}^{k} \alpha_{spj} f_j(x_{ip})) + \beta \sum_{p=1}^{l} \sum_{j=1}^{k} \alpha_{spj}^2) \to \min_{\vec{\alpha}_1, \ldots, \vec{\alpha}_m} \ ,$$

где y_{si} — класс i-го пептида ($y_{si} = 1$, если s-ая протеаза разрезает i-ый пептид, иначе $y_{si} = -1$), $\sigma(z) = 1/(1 + exp(-z))$ — логистическая (сигмоидная) функция.

Оптимизация этих функционалов может быть осуществлена любым итерационным алгоритмом. В работе применялся один из квази-ньютоновских методов, основанный на аппроксимации обратной матрицы вторых производных последовательностью матриц, при расчете которых использовались приращения первых производных и оптимизируемых параметров на предыдущем шаге. Шаг алгоритма подбирался методом дробления шага [38].

Композиция алгоритмов классификации и регрессии

Для решения задачи ранжирования пептидов помимо моделей, построенных для задач предсказания и классификации, использовалась их композиция. Для этого обе модели обучались по отдельности (регрессионная — на множестве пептидов с известным значением эффективности, классификации — на объединении этих пептидов, помеченных классом $+1$, со случайными, помеченными классом -1).

При ранжировании сначала к пептиду применялся классификатор: вычислялась соответствующая линейная функция. Если ее значение было меньше нуля, то пептиду приписывалось это значение. Если значение было больше нуля, то применялась регрессионная модель и пептиду приписывается предсказанное ею значение. Далее, пептиды упорядочивались по приписанным значениям.

Оптимизация кусочно-квадратичной функции потерь

Для дополнения регрессионной модели информацией о низкой эффективности протеолиза случайных пептидов можно добавить их в обучающую выборку и оптимизировать кусочно-квадратичную функцию потерь. Для пептидов с известным значением эффективности использовалась квадратичная функция потерь. Для случайных пептидов ошибка полагалась равной нулю, если прогнозируемое значение ниже минимального значения эффективности разрезания всех имеющихся экспериментальных пептидов для данной протеазы, и равной квадрату отклонения прогноза от минимального значения, если прогнозируемое значение выше.

Пусть пептиды задаются k-мерными векторами числовых признаков $x = (x_1, \ldots, x_k)$. Оптимизированный функционал будет иметь вид:

$$L(\vec{\alpha}) = \sum_{i=1}^{n} (y_i - \langle \vec{\alpha}, \vec{x}_i \rangle)^2 + \sum_{i=1}^{m} [y_{min} - \langle \vec{\alpha}, \vec{x}_i \rangle]^2 + \beta \sum_{i=1}^{k} \alpha_i^2 \to \min_{\vec{\alpha}} \ ,$$

где n — число пептидов обучающей выборки с известным значением активности, m — число случайных пептидов, y_{min} — минимальное значение эффективности в выборке, $\vec{\alpha}$ - вектор коэффициентов, $\vec{x}_i$ — вектор признаков i-го пептида, $\beta \sum_{i=1}^{k} \alpha_i^2$ — регуляризующее слагаемое, β — априори зафиксированный параметр регуляризации, $[x] = x$, при $x < 0$, иначе $[x] = 0$.

Эта функция является кусочно-квадратичной, так как в интервалах, где знаки всех выражений в квадратных скобках постоянны, она является суммой квадратичных функций. Однако, такая функция потерь не является непрерывно дифференцируемой и имеет в общем случае более одного локального минимума, что сильно затрудняет оптимизацию. В настоящей работе применялся следующий эвристический метод. Всем случайным пептидам приписывалось минимальное значение эффективности и обучалась гребневая регрессионная модель. Далее, все случайные пептиды, для которых прогнозируемое полученной моделью значение было ниже минимального, исключались из выборки. Гребневая регрессия заново обучалась на оставшихся пептидах. Процесс повторялся до тех пор, пока обучающаяся выборка не стабилизировалась.

Эксперименты показали, что процесс сходится довольно быстро и конечная обучающая выборка не совпадает с множеством пептидов с известными значениями эффективности, т.е. удается использовать дополнительную информацию при обучении регрессионной модели.

Жадный отбор признаков

Отбор признаков состоит в переборе подмножеств признаков с обучением алгоритма на части обучающей выборки и оценкой ошибки на оставшейся части. Выбирается множество, дающее наименьшую ошибку.

На практике, перебрать все подмножества нельзя. Поэтому использовался жадный алгоритм: к текущему наилучшему множеству признаков последовательно добавлялся каждый из оставшихся. Выбирался признак, добавление которого давало наилучший классификатор (регрессионную модель). Этот признак добавлялся к наилучшему множеству и процесс повторялся.

3. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

3.1. Определение релевантных позиций пептидов

Отбор релевантных позиций проводился жадным алгоритмом отдельно для задачи предсказания эффективности разрезания пептидов и задачи классификации пептидов на основе выравнивания длиной 10 а.о. На каждой итерации по текущему отобранному множеству обучались модели (линейную регрессию для задачи предсказания эффективности и логистическую регрессию для задачи классификации) и оценивалось качество прогноза. В обоих случаях процесс повторялся для разных разбиений выборки при скользящем контроле, и результаты усреднялись по всем разбиениям.

По результатам тестов были построены графики зависимости точности прогноза от числа отобранных позиций, а также таблицы, показывающие, какие позиции были отобраны на каждой итерации отбора. Для разных разбиений выборки могут быть отобраны различные наборы позиций. Поэтому приводится только частота появления данной позиции во множестве отобранных на некотором шаге алгоритма позиций, т.е. частота появления позиции в отобранных наборах длины от 1 до 10. Позиции упорядочены по суммарной частоте (сумме частот в наборах всевозможных длин).

На рис. 2 представлена зависимость ошибки прогноза гребневой регрессии для всех ММР от количества позиций, отобранных жадным алгоритмом. Видно, что ошибка быстро убывает, пока число позиций не достигает пяти-шести, после чего ошибка либо убывает медленно, либо медленно растет. Можно сделать вывод, что для предсказания эффективности разумно использовать 6 позиций.

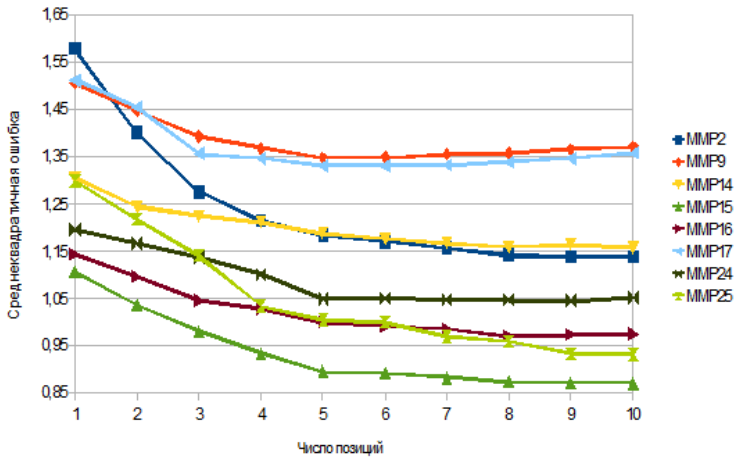

Рис. 2. Зависимость средней ошибки прогноза гребневой регрессии от количества позиций, отобранных жадным алгоритмом.

В таблице 1 представлены частоты позиций $P_5, \ldots, P_1, P'_1, \ldots, P'_5$ в наборах различной длины $N = \{1, \ldots, 10\}$, отобранных жадным алгоритмом на основе гребневой регрессии в задаче предсказания эффективности разрезания пептидов для всех изученных ММР. В большинстве случаев стабильно отбираются шесть позиций $P_3, \ldots, P_1, P'_1, \ldots, P'_3$, что согласуется с видом рис. 2. Порядок этих позиций по суммарной частоте у всех рассмотренных протеаз, кроме ММР17, отличается лишь взаимным положением P_2 и P'_2 на четвертом и пятом шаге: в случае ММР 2,9,14,16 — P_2 предшествует P'_2, в случае ММР 15, 24, 25 — наоборот, причем в

обоих случаях суммарная частота их появления на четвертом шаге равна единице, т.е. всегда выбирается одна из них.

Таблица 1. Частоты позиций $P_5, \ldots, P_1, P'_1, \ldots, P'_5$ в наборах различной длины $N = \{1, \ldots, 10\}$, отобранных жадным алгоритмом на основе гребневой регрессии в задаче предсказания эффективности разрезания пептидов для всех изученных ММР.

N	MMP2										MMP15									
	P'_1	P_3	P_1	P_2	P'_2	P'_3	P_4	P_5	P'_4	P'_5	P'_1	P_3	P_1	P'_2	P_2	P'_3	P'_4	P_4	P_5	P'_5
1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
2	1	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0
3	1	1	0.91	0.09	0	0	0	0	0	0	1	1	0.99	0.01	0	0	0	0	0	0
4	1	1	0.99	0.74	0.28	0	0	0	0	0	1	1	1	0.71	0.29	0	0	0	0	0
5	1	1	1	1	0.94	0.05	0	0.01	0	0	1	1	1	0.86	0.99	0.15	0	0	0	0
6	1	1	1	1	1	0.79	0.1	0.11	0	0	1	1	1	1	1	0.86	0.09	0.03	0.03	0
7	1	1	1	1	1	0.96	0.46	0.39	0.14	0.05	1	1	1	1	1	0.96	0.45	0.29	0.23	0.08
8	1	1	1	1	1	1	0.73	0.58	0.43	0.28	1	1	1	1	1	1	0.75	0.53	0.41	0.31
9	1	1	1	1	1	1	0.83	0.75	0.78	0.65	1	1	1	1	1	1	0.94	0.84	0.65	0.58
10	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
N	MMP9										MMP16									
	P'_1	P_3	P_1	P_2	P'_2	P'_3	P_5	P_4	P'_4	P'_5	P'_1	P_3	P_1	P_2	P'_2	P'_3	P_5	P'_4	P_4	P'_5
1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
2	1	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0
3	1	1	0.81	0.18	0.01	0	0	0	0	0	1	1	0.98	0	0.03	0	0	0	0	0
4	1	1	0.93	0.65	0.43	0	0	0	0	0	1	1	1	0.63	0.36	0.01	0	0	0	0
5	1	1	1	1	0.99	0.01	0	0.01	0	0	1	1	1	0.9	0.99	0.1	0	0.01	0	0
6	1	1	1	1	1	0.51	0.29	0.2	0	0	1	1	1	0.99	1	0.86	0.01	0.1	0.04	0
7	1	1	1	1	1	0.89	0.58	0.39	0.13	0.03	1	1	1	1	1	0.99	0.4	0.33	0.13	0.16
8	1	1	1	1	1	0.99	0.75	0.54	0.58	0.15	1	1	1	1	1	0.99	0.63	0.61	0.46	0.31
9	1	1	1	1	1	1	0.94	0.71	0.86	0.49	1	1	1	1	1	1	0.86	0.79	0.88	0.48
10	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
N	MMP14										MMP17									
	P'_1	P_3	P_1	P_2	P'_2	P'_3	P_5	P_4	P'_4	P'_5	P'_1	P_3	P_2	P_1	P'_2	P'_3	P_4	P'_4	P'_5	P_5
1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
2	1	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0
3	1	1	0.9	0.06	0.04	0	0	0	0	0	1	1	0.38	0.38	0.25	0	0	0	0	0
4	1	1	1	0.6	0.4	0	0	0	0	0	1	1	0.6	0.55	0.76	0.09	0	0	0	0
5	1	1	1	1	0.96	0.01	0.03	0	0	0	1	1	0.9	0.91	0.83	0.34	0	0.03	0	0
6	1	1	1	1	1	0.6	0.25	0.11	0.04	0	1	1	0.99	1	1	0.88	0.09	0.05	0	0
7	1	1	1	1	1	0.88	0.58	0.29	0.18	0.09	1	1	1	1	1	0.98	0.51	0.38	0.06	0.08
8	1	1	1	1	1	0.99	0.79	0.59	0.45	0.19	1	1	1	1	1	0.99	0.81	0.58	0.36	0.26
9	1	1	1	1	1	1	0.93	0.84	0.81	0.43	1	1	1	1	1	1	0.91	0.86	0.79	0.44
10	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
N	MMP24										MMP25									
	P'_1	P_3	P_1	P'_2	P_2	P'_3	P_5	P_4	P'_4	P'_5	P'_1	P_3	P_1	P'_2	P_2	P'_3	P'_4	P_4	P_5	P'_5
1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
2	1	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0
3	1	1	0.96	0.01	0.03	0	0	0	0	0	1	1	0.96	0.04	0	0	0	0	0	0
4	1	1	0.96	0.84	0.2	0	0	0	0	0	1	1	0.98	0.86	0.15	0.01	0	0	0	0
5	1	1	1	0.98	0.99	0.03	0.01	0	0	0	1	1	1	0.93	0.96	0.11	0	0.03	0	0
6	1	1	1	0.99	1	0.58	0.24	0.19	0.01	0	1	1	1	0.99	1	0.93	0.01	0.03	0.05	0
7	1	1	1	1	1	0.93	0.49	0.43	0.16	0	1	1	1	1	1	0.99	0.31	0.4	0.26	0.04
8	1	1	1	1	1	1	0.73	0.61	0.56	0.1	1	1	1	1	1	1	0.68	0.66	0.46	0.2
9	1	1	1	1	1	1	0.88	0.86	0.91	0.35	1	1	1	1	1	1	0.94	0.84	0.69	0.54
10	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

На рис. 3 представлена зависимость ошибки классификации логистической регрессии для всех ММР от количества позиций, отобранных жадным алгоритмом. Качественно график совпадает с графиком гребневой регрессии.

В таблице 2 представлены частоты позиций $P_5, \ldots, P_1, P'_1, \ldots, P'_5$ в наборах различной длины $N = \{1, \ldots, 10\}$, отобранных жадным алгоритмом на основе логистической регрессии в задаче классификации пептидов для всех изученных ММР. Как и в случае с гребневой регрессией, стабильно отбираются шесть позиций $P_3, \ldots, P_1, P'_1, \ldots, P'_3$. Порядок этих позиций по суммарной частоте одинаков у всех рассмотренных протеаз, кроме ММР14, у которой P_1 выбирается чаще P'_2, причем суммарная частота их появления на третьем шаге близка к единице. Также близка к единице суммарная частота появления P_2 и P'_3 на пятом шаге, т.е. как правило выбирается одни из них.

Суммарная частота выбора позиции при отборе характеризует относительную важность позиции для моделирования первичной специфичности. Важность позиций можно оценить и

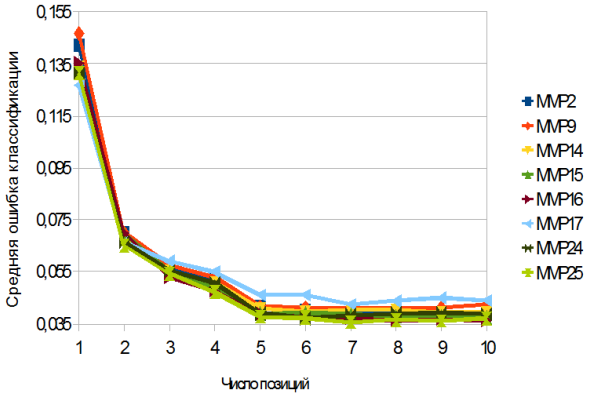

Рис. 3. Зависимость средней ошибки прогноза логистической регрессии от количества позиций, отобранных жадным алгоритмом.

Таблица 2. Частоты позиций $P_5, \ldots, P_1, P'_1, \ldots, P'_5$ в наборах различной длины $N = \{1, \ldots, 10\}$, отобранных жадным алгоритмом на основе логистической регрессии в задаче классификации пептидов для всех изученных ММР.

N	MMP2										MMP15									
	P'_1	P_3	P'_2	P_1	P_2	P'_3	P_5	P_4	P'_5	P'_4	P'_1	P_3	P'_2	P_1	P_2	P'_3	P_5	P_4	P'_5	P'_4
1	0.99	0.01	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
2	1	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0
3	1	1	0.86	0.06	0	0	0.03	0	0.05	0	1	1	0.98	0.01	0	0	0.01	0	0	0
4	1	1	0.95	0.91	0.03	0	0.06	0	0.05	0	1	1	1	0.99	0	0	0.01	0	0	0
5	1	1	1	0.98	0.7	0.18	0.09	0	0.06	0	1	1	1	1	0.73	0.26	0.01	0	0	0
6	1	1	1	1	0.93	0.55	0.21	0.24	0.08	0	1	1	1	1	0.99	0.85	0.16	0	0	0
7	1	1	1	1	0.98	0.86	0.44	0.63	0.1	0	1	1	1	1	1	0.9	0.56	0.28	0.19	0.08
8	1	1	1	1	1	1	1	0.9	0.1	0	1	1	1	1	1	1	1	0.98	0.03	0
9	1	1	1	1	1	1	1	0.99	0.5	0.51	1	1	1	1	1	1	1	1	0.84	0.16
10	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

N	MMP9										MMP16									
	P'_1	P_3	P'_2	P_1	P_2	P'_3	P_5	P_4	P'_5	P'_4	P'_1	P_3	P'_2	P_1	P_2	P'_3	P_5	P_4	P'_5	P'_4
1	0.56	0.44	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
2	1	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0
3	1	1	0.89	0.03	0.03	0	0.04	0	0.03	0	1	1	0.94	0.06	0	0	0	0	0	0
4	1	1	0.96	0.9	0.06	0	0.05	0	0.03	0	1	1	1	0.98	0	0	0.03	0	0	0
5	1	1	1	1	0.86	0.06	0.05	0	0.03	0	1	1	1	1	0.5	0.48	0.03	0	0	0
6	1	1	1	1	0.99	0.73	0.21	0.05	0.03	0	1	1	1	1	0.93	0.89	0.18	0.01	0	0
7	1	1	1	1	1	0.98	0.78	0.23	0.03	0	1	1	1	1	1	0.99	0.85	0.09	0.08	0
8	1	1	1	1	1	1	1	0.98	0.03	0	1	1	1	1	1	1	0.99	0.74	0.28	0
9	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0.83	0.18
10	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

N	MMP14										MMP17									
	P'_1	P_3	P_1	P'_2	P_2	P'_3	P_4	P_5	P'_5	P'_4	P'_1	P_3	P'_2	P_1	P_2	P'_3	P_5	P_4	P'_5	P'_4
1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
2	1	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0
3	1	1	0.68	0.28	0.04	0	0	0	0.01	0	1	1	0.91	0.08	0	0	0.01	0	0	0
4	1	1	0.98	0.98	0.04	0	0	0	0.01	0	1	1	1	0.95	0	0	0.05	0	0	0
5	1	1	1	1	0.89	0.09	0	0	0.03	0	1	1	1	1	0.46	0.31	0.23	0	0	0
6	1	1	1	1	1	0.83	0.08	0.04	0.06	0	1	1	1	1	0.86	0.91	0.23	0	0	0
7	1	1	1	1	1	0.96	0.73	0.18	0.13	0.01	1	1	1	1	1	1	0.63	0.36	0.01	0
8	1	1	1	1	1	1	0.86	0.99	0.14	0.01	1	1	1	1	1	1	1	0.99	0.01	0
9	1	1	1	1	1	1	1	1	0.85	0.15	1	1	1	1	1	1	1	1	0.69	0.31
10	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

N	MMP24										MMP25									
	P'_1	P_3	P'_2	P_1	P_2	P'_3	P_5	P'_5	P_4	P'_4	P'_1	P_3	P'_2	P_1	P_2	P'_3	P_5	P_4	P'_5	P'_4
1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
2	1	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0
3	1	1	0.71	0.1	0.06	0	0.01	0.11	0	0	1	1	0.84	0.14	0	0	0.03	0	0	0
4	1	1	0.91	0.83	0.14	0	0.01	0.11	0	0	1	1	1	0.98	0	0	0.03	0	0	0
5	1	1	1	0.95	0.91	0.01	0.01	0.11	0	0	1	1	1	1	0.5	0.4	0.1	0	0	0
6	1	1	1	1	1	0.53	0.13	0.26	0.09	0	1	1	1	1	0.91	0.94	0.15	0	0	0
7	1	1	1	1	1	0.85	0.51	0.34	0.3	0	1	1	1	1	1	1	0.39	0.6	0.01	0
8	1	1	1	1	1	1	1	0.38	0.63	0	1	1	1	1	1	1	1	0.93	0.08	0
9	1	1	1	1	1	1	1	0.96	0.99	0.05	1	1	1	1	1	1	1	1	0.89	0.11
10	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

по значениям коэфициентов при соответствующих признаках в построенных моделях. В таблице 3 представлены численные оценки вклада позиций $P_3, \ldots, P_1, P_1', \ldots, P_3'$ в регрессионные и классификационные модели для всех изученных протеаз. Для численной оценки вклада позиции в модель данной протеазы модули коэффициентов всех аминокислотных остатков в данной позиции усреднялись. После этого средние значения всех позиций для данной протеазы складывались и делились на полученную сумму. Как видно из таблицы, наибольший вклад во все модели для всех протеаз вносит позиция P_1', на втором месте — позиция P_3, что согласуется с частотами их выбора при отборе.

Таблица 3. Численные оценки вклада позиций $P_3, \ldots, P_1, P_1', \ldots, P_3'$ в регрессионные и классификационные модели для всех изученных протеаз. Представлены алгоритмы гребневой регрессии с параметризацией аминокислот (RR (п.а.)), логистической регрессии с параметризацией аминокислот (LR (п.а.)) и метода оптимизации кусочно-квадратичной функции потерь (PQ).

Протеаза	RR (п.а.)						LR (п.а.)						PQ					
	P_3	P_2	P_1	P_1'	P_2'	P_3'	P_3	P_2	P_1	P_1'	P_2'	P_3'	P_3	P_2	P_1	P_1'	P_2'	P_3'
MMP2	0.17	0.15	0.15	0.27	0.14	0.12	0.18	0.17	0.11	0.21	0.16	0.18	0.24	0.12	0.11	0.24	0.13	0.16
MMP9	0.16	0.15	0.1	0.41	0.1	0.09	0.17	0.16	0.11	0.20	0.17	0.17	0.27	0.13	0.10	0.24	0.12	0.14
MMP14	0.14	0.16	0.22	0.28	0.14	0.06	0.17	0.15	0.12	0.21	0.17	0.18	0.20	0.12	0.14	0.25	0.12	0.16
MMP15	0.11	0.10	0.25	0.31	0.16	0.07	0.18	0.15	0.12	0.22	0.15	0.18	0.20	0.10	0.10	0.28	0.15	0.16
MMP16	0.11	0.11	0.23	0.31	0.19	0.06	0.18	0.14	0.12	0.22	0.16	0.18	0.20	0.11	0.11	0.27	0.14	0.16
MMP17	0.26	0.10	0.06	0.40	0.14	0.04	0.22	0.13	0.10	0.25	0.15	0.17	0.26	0.09	0.11	0.30	0.12	0.12
MMP24	0.12	0.15	0.26	0.26	0.16	0.06	0.18	0.15	0.12	0.23	0.15	0.18	0.21	0.10	0.12	0.28	0.14	0.15
MMP25	0.20	0.16	0.12	0.28	0.14	0.10	0.20	0.13	0.11	0.23	0.15	0.18	0.23	0.08	0.11	0.29	0.15	0.14

3.2. Сравнение использованных моделей

В табл. 4 представлены кросс-валидационные оценки качества различных моделей, обученных на пептидах длиной 6 а.о., при различных постановках задачи. Приведены значения соответствующих функционалов ошибки, усредненные по всем протеазам. В задаче предсказания активности протеаз линейная регрессия показала существенно лучший результат, чем метод k ближайших соседей. Это можно объяснить тем, что при расчете меры близости все позиции считались равнозначными, в то время как линейная регрессия учитывает разную степень значимости позиций. Тот же эффект наблюдается в задаче классификации: логистическая регрессия и наивный байесовский классификатор значительно опережают метод ближайших соседей.

Таблица 4. Оценка качества различных моделей, обученных на пептидах длины 6 а.о., при различных постановках задачи. RR — гребневая регрессия, LR — логистическая регрессия, PQ — оптимизация кусочно-квадратичной функции потерь, kNN — метод k ближайших соседей, NB — наивный байесовский классификатор. В алгоритмах, помеченных (п.а.), применялась параметризация аминокислот. Указаны средние ошибки (AE) и среднеквадратичные отклонения (SD).

Алгоритм	Задача					
	Регрессия		Классификация		Ранжирование	
	AE	SD	AE	SD	AE	SD
RR	1,1148	0,0051			0,2019	0,0011
RR (п.а.)	1,1099	0,0049			0,1676	0,0016
kNN	1,2533	0,0052	0,0394	0,0006	0,1336	0,0009
NB			0,0299	0,0007	0,1856	0,0008
LR			0,0276	0,0006	0,1675	0,0007
LR (п.а.)			0,0260	0,0006	0,1677	0,0007
PQ					0,1405	0,0006
Композиция					0,1268	0,0005

Важно также, что ошибка логистической регрессии ниже, чем у наивного байесовского классификатора: позиции пептида не являются независимыми.

Эксперименты показали, что использование параметризации аминокислот повышает качество моделей, как в задаче прогнозирования активности, так и в задаче классификации.

Задача ранжирования позволяет сравнить все модели одновременно. Наиболее эффективной моделью оказалась композиция алгоритмов. Видно также, что оптимизация кусочно-квадратичной функции потерь дает лучшие результаты на задаче ранжирования, по сравнению с классификаторами. Хорошие результаты при ранжировании показал метод k ближайших соседей, оказавшийся вторым по качеству.

В базе экспериментально зафиксированных событий протеолиза CutDB имеются данные для трех из исследованных протеаз: MMP2, MMP9 и MMP14. На рис. 4, 5, 6 представлены ROC-кривые для различных алгоритмов. Для всех трех протеаз лучшим по интегральной характеристике (AUC) оказался метод k ближайших соседей. Однако его преимущество проявляется только при очень больших значениях чувствительности (и невысокой специфичности). При меньших значениях чувствительности лидируют алгоритмы классификации: логистическая регрессия (обычная и с параметризацией аминокислот). Так как на CutDB решалась задача классификации, то этого следовало ожидать. Алгоритмы ранжирования показали схожую эффективность.

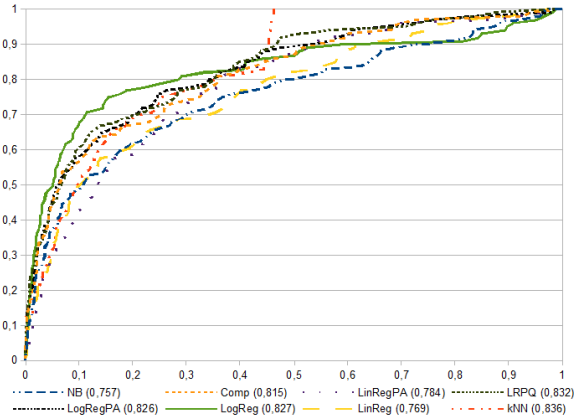

Рис. 4. ROC-кривые различных алгоритмов для протеазы MMP2, построенные по CutDB. В скобках в легенде указано значение AUC для данного алгоритма. Обозначения алгоритмов: NB — наивный байесовский классификатор, kNN — метод k ближайших соседей, LinReg — многомерная линейная регрессия, LogReg — логистическая регрессия, LogRegPA — логистическая регрессия с параметризацией аминокислот, LinRegPA — линейная регрессия с параметризацией аминокислот, Comp — композиция LogRegPA и LinRegPA, LRPQ — оптимизация кусочно-квадратичной функции потерь.

Абсолютные значения точности на CutDB существенно ниже результатов, полученных для алгоритмов классификации с помощью кросс-валидации. Это можно объяснить двумя причинами. Во-первых, для разрезания реального белка протеазе необходим физический доступ к сайту разрезания [39]. Многие подходящие сайты могут иметь неподходящую вторичную структуру или могут быть погружены в белковую глобулу, и поэтому быть недоступны. Такие пептиды будут помечены как отрицательные и будут засчитаны как ложно положительные ответы. Во-вторых, в базе CutDB содержаться данные различных экспериментов, которые могли быть проведены в различных условиях. Важным параметром является время, предоставленное протеазам для разрезания исследуемых белков. Протеазы могут просто не успеть разрезать сайт, обладающий средним сродством, либо, наоборот, успеть разрезать даже сайт с низким сродством. Это обстоятельство может приводить к росту числа как ложно положительных, так и ложно отрицательных ответов.

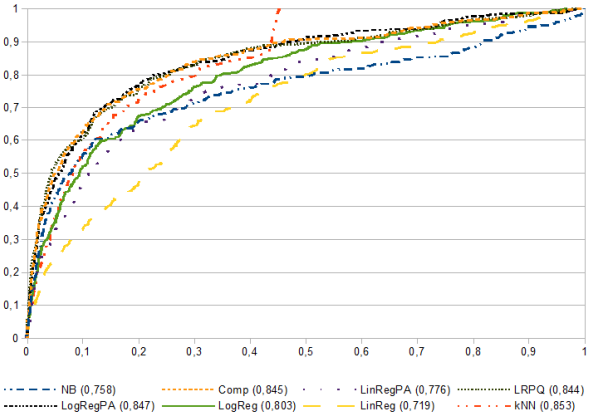

Рис. 5. ROC-кривые различных алгоритмов для протеазы MMP9, построенные по CutDB. В скобках в легенде указано значение AUC для данного алгоритма. Обозначения алгоритмов — см. рис. 4.

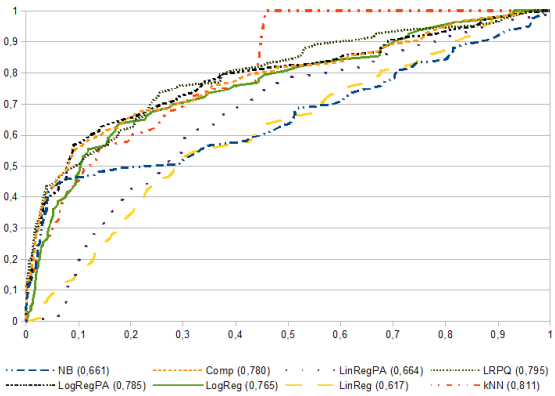

Рис. 6. ROC-кривые различных алгоритмов для протеазы MMP14, построенные по CutDB. В скобках в легенде указано значение AUC для данного алгоритма. Обозначения алгоритмов — см. рис. 4.

4. ЗАКЛЮЧЕНИЕ

Были построены регрессионные и классификационные модели, а также модели ранжирования, позволяющие эффективно решать соответствующие задачи, что позволяет надеяться на их эффективное использование для предсказания белковых субстратов исследованных протеаз.

При построении моделей использовался разработанный метод сокращения размерности пространства признаков на основе параметризации аминокислот, повышающий качество гребневой и логистической регрессии в задачах предсказания эффективности, классификации и ранжирования пептидов.

Реализован и протестирован метод обучения линейной регрессионной модели с помощью оптимизации кусочно-квадратичной функции потерь. Вычислительные эксперименты показали, что одновременное использование значений эффективности и случайных пептидов позволяет получить более точные модели.

5. БЛАГОДАРНОСТИ

Благодарим лаборатории Джефа Смита и Андрея Остермана института медицинских исследований Сэнфорд-Бернэм, Ла Хойя, США за предоставленные данные.

Исследование выполнено при финансовой поддержке федеральной целевой программы "Научные и научно-педагогические кадры инновационной России"в рамках научного проекта "Реконструкция сигнальных и протеолитичеких взаимодействий белок-белок и белок-лиганд методами структурной биоинформатики и машинного обучения"заявка №2012-1.2.2-12-000-1013-079.

СПИСОК ЛИТЕРАТУРЫ

1. Gross, J., Lapiere, C. Collagenolytyc activity in amphibian tissues: a tissue culture assay. *Proc Natl Acad Sci USA*, 1962, vol. 48, no. 6, pp. 1014?22.

2. Eisen, A., Jeffrey, J., Gross, J. Human skin collagenase. Isolation and mechanism of attack on the collagen molecule. *Biochim Biophys Acta*, 1968, vol. 151, no. 3, pp. 637?45.

3. Parks, W.C., Wilson, C.L., Lopez-Boado, Y.S. Matrix metalloproteinases as modulators of inflammation and innate immunity. *Nat. Rev. Immunol.*, 2004, pp. 617-29

4. Van Lint, P., Libert, C. Chemokine and cytokine processing by matrix metalloproteinases and its effect on leukocyte migration and inflammation. *J. Leukoc. Biol.*, 2007, vol. 82, no. 6, pp. 1375-81

5. Rodriguez, D., Morrison, C.J., Overall, C.M. Matrix metalloproteinases: what do they not do? New substrates and biological roles identified by murine models and proteomics. *Biochim. Biophys. Acta.*, 2010, vol. 1803, no. 1, pp. 39-54

6. Zucker, S., Cao, J., Chen, W.-T. Critical appraisal of the use of matrix metalloproteinase inhibitors in cancer treatment. *Nature*, 2000, vol. 19, n. 56, pp. 6642-6650.

7. Yip, D., Ahmad, A., Karapetis, C.S., Hawkins, C.A., Harper, P.G. Matrix metalloproteinase inhibitors: applications in oncology. *Invest New Drugs*, 1999, vol. 17(4), pp. 387-99.

8. Liu, P., Sun, M., Sader, S. Matrix metalloproteinases in cardiovascular disease. *C. J. Cardiol.*, 2006, vol. 22, pp. 25-30.

9. Lovejoy, B., Cleasby, A., Hassell, A.M., Longley, K., Luther, M.A., Weigl, D., et al. Structure of the catalytic domain of fibroblast collagenase complexed with an inhibitor. *Science*, 1994, vol. 263, pp. 375-377.

10. Bode, W., Reinemer, P., Huber, R., Kleine, T., Schnierer, S., Tschesche, H. The X-ray crystal structure of the catalytic domain of human neutrophil collagenase inhibited by a substrate analogue reveals the essentials for catalysis and specificity. *EMBO J*, 1994, vol. 13, pp. 1263-1269.

11. Borkakoti, N., Winkler, F.K., Williams, D.H., D?Arcy, A., Broadhurst, M.J., Brown, P.A., et al. Structure of the catalytic domain of human fibroblast collagenase complexed with an inhibitor. *Nat Struct Biol*, 1994, vol. 1, pp. 106-110.

12. Stams, T., Spurlino, J.C., Smith, D.L., Wahl, R.C., Ho, T.F., Qoronfleh, M.W., et al. Structure of human neutrophil collagenase reveals large S1' specificity pocket. *Nat Struct Biol*, 1994, vol. 1, pp. 119-123.

13. Becker, J.W., Marcy, A.I., Rokosz, L.L., Axel, M.G., Burbaum, J.J., Fitzgerald, P.M., et al. Stromelysin-1: three-dimensional structure of the inhibited catalytic domain and of the C-truncated proenzyme. *Protein Sci*, 1995, vol. 4, pp. 1966-1976.

14. Li, J., Brick, P., O'Hare, M.C., Skarzynski, T., Lloyd, L.F., Curry, V.A., et al. Structure of full-length porcine synovial collagenase reveals a C-terminal domain containing a calcium-linked, four-bladed beta-propeller. Structure, 1995, vol. 3, pp. 541-549.

15. Bode, W., Gomis-Ruth, F.-X., Stockler, W. Astacins, serralysins, snake venom and matrix metalloproteinases exhibit identical zinc-binding environments (HEXXHXXGXXH and Met-turn) and topologies and should be grouped into a common family, the metzincins. *FEBS Lett.*,1993, vol. 331, pp. 134-140.

16. Windsor, L. J., Bodden, M. K., Birkedal-Hansen, B., Engler, J. A., Birkedal-Hansen, H. Mutational analysis of residues in and around the active site of human fibroblast-type collagenase. *J. Biol. Chem.*, 1994, vol. 269, pp. 26201-26207

17. Pourmotabbed, T., Solomon, T. L., Hasty, K. A., Mainardi, C. L. Characteristics of 92 kDa type IV collagenase/gelatinase produced by granulocytic leukemia cells: structure, expression of cDNA in E. coli and enzymic properties. *Biochim. Biophys. Acta.*, 1994, vol. 1204, pp. 97-107.

18. Bode, W., Fernandez-Catalan C., Tschesche, H., Grams, F., Nagase, H., Maskos, K. Structural properties of matrix metalloproteinases. *Cell Mol. Life Sci.*, 1999, vol. 55, pp. 639-652.

19. Schechter, I., Berger, A. On the size of the active site in proteases. I. Papain. *Biochem. Biophys. Res. Commun.*, 1967, vol. 27, pp. 157-162

20. Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M.R., Appel, R.D., Bairoch, A. Protein Identification and Analysis Tools on the ExPASy Server. In: *The Proteomics Protocols Handbook* Edited by Walker JM. Humana Press, 2005, pp. 571-607.

21. Lohmuller, T., Wenzler, D., Hagemann, S., Kiess, W., Peters, C., Dandekar, T., Reinheckel, T. Toward computer-based cleavage site prediction of cysteine endopeptidases. *Biol. Chem.*, 2003, vol. 384, pp. 899-909.

22. Bredemeyer, A. J., Lewis, R. M., Malone, J. P., Davis, A. E., Gross, J., Townsend, R. and Ley, T. J. A proteomic approach for the discovery of protease substrates. *Proc. Natl. Acad. Sci. USA*, 2004, vol. 101, no. 32, pp. 11785-11790.

23. Tam, E. M., Morrison, C. J., Wu, Y. I., Stack, M. S., Overall, C. M. Membrane protease proteomics: Isotope-coded affinity tag MS identification of undescribed MT1-matrix metalloproteinase substrates. *Proc. Natl. Acad. Sci. USA*, 2004, vol. 101, no. 18, pp. 6917-6922.

24. Tompa, P., Buzder-Lantos, P., Tantos, A., Farkas, A., Szilagyi, A., Banoczi, Z., Hudecz, F., Friedrich, P. On the sequential determinants of calpain cleavage. *J. Biol. Chem.*, 2004, vol. 279, pp. 20775-20785.

25. Garay-Malpartida, H.M., Occhiucci, J.M., Alves, J., Belizario, J.E. CaSPredictor: a new computer-based tool for caspase substrate prediction. *Bioinformatics*, 2005, vol. 21 (Suppl 1): pp. i169-i176.

26. Backes, C., Kuentzer, J., Lenhof, H.-P., Comtesse, N., Meese, E. GraBCas: a bioinformatics tool for score-based prediction of Caspase- and Granzyme B-cleavage sites in protein sequences. *Nucleic Acids Research*, 2005, vol. 33, pp. W208-W213.

27. Boyd, S.E., Pike, R.N., Rudy, G.B., Whisstock, J.C., Garcia de la Banda, M. PoPS: a computational tool for modeling and predicting protease specificity. *J. Bioinform. Comput. Biol.*, 2005, vol. 3, pp. 551-585.

28. Yang, Z.R. Prediction of caspase cleavage sites using Bayesian bio-basis function neural networks. *Bioinformatics*, 2005, vol. 21, pp. 1831-1837.

29. Wee, L.J., Tan, T.W., Ranganathan, S. SVM-based prediction of caspase substrate cleavage sites. *BMC Bioinformatics*, 2006, vol. 7 (Suppl. 5), S14.

30. Rogers, S., Wells, R., Rechsteiner, M. Amino acid sequences common to rapidly degraded proteins: the PEST hypothesis. *Science*, 1986, vol. 234, pp. 364-368.

31. Rechsteiner, M., Rogers, S. PEST sequences and regulation by proteolysis. *Trends Biochem Sci*, 1996, vol. 21, pp. 267-271.

32. Ratnikov, B., Cieplak, P., Smith, J.W. High Throughput Substrate Phage Display for Protease Profiling. *Methods in Molecular Biology*, 2009, vol. 539, pp. 1-22.

33. Igarashi, Y., Eroshkin, A., Gramatikova, S., Gramatikoff, G., Zhang, Y., Smith, J.W., Osterman, A.L., Godzik, A. CutDB: a proteolytic event database. *Nucleic Acids Res.*, 2007, vol. 35 (Database issue), pp. D546-9. Epub 2006 Nov 16.

34. Domingos, P., Pazzani, M. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 1997, vol. 29, pp. 103-137.

35. Rao, C.R. *Linear statistical inference and it's applications (2nd ed.)*, John Wiley & Sons, 1973, New York.

36. Hosmer, D., Lemeshow, S. *Applied Logistic Regression, 2nd ed.*, 2000, Wiley, New York, Chichester.

37. Peng, H.C., Long, F., Ding, C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, 27, pp. 1226-1238.

38. Fletcher, R. *Practical methods of optimization (2nd ed.)*, John Wiley & Sons, 1987, New York.

39. Kazanov, M.D., Igarashi, Y., Eroshkin, A.M., Cieplak, P., Ratnikov, B., Zhang, Y., Li, Z., Godzik, A., Osterman, A.L., Smith, J.W. Structural determinants of limited proteolysis. *J. Proteome Res.*, 2011, vol. 10, no. 8, pp. 3642-51.

MMP proteases family primary specificity prediction using machine learning methods

G.G. Fedonin, M.D. Kazanov

Eight proteases from different subgroups of MMP family were studied. Predictive models of primary specificity were built for these proteases based on proteolytic activity measurements for 1369 peptides, which were selected experimentally from 64 million of peptides of length six using phage display and were known to be actively cleaved by at least one of considered proteases.

Linear regression models and kNN were used to predict proteolytic activity of each protease on peptide given it's amino acid sequence. Naive Bayes classifier, logistic regression and kNN were used to solve binary classification problem, in which peptides with known activity were considered as positive examples while random peptides were considered as negative ones.

Greedy forward feature selection was used to determine amino acid sequence positions of peptides, which are relevant for specificity prediction.

A method of dimensionality reduction based on parametrization of amino acids was suggested and was shown to improve prediction quality both for regression and classification.

Special training algorithm for linear regression model was suggested: random peptides were added to training set. Training error for random peptide was defined to be zero, if predicted value is less than minimal activity in the sample for given protease, and, otherwise, to be equal to squared deviation of predicted value from minimal activity (same as in OLS algorithm).

Models' validation was performed on CutDB — a database of experimentally observed proteolytic events. ROC curves were built using CutDB data to demonstrate efficiency of suggested models.

KEYWORDS: bioinformatics, proteases, matrix metalloproteinases, primary specificity of proteases, machine learning, feature space dimensionality reduction.