Математическая биология и биоинформатика
2017. Т. 12. № 2. С. 435–445. doi: 10.17537/2017.12.435

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

УДК: 57.087

Анализ многомерных данных пептидных микрочипов с использованием метода проекции на латентные структуры

Анисимов Д.С.[*1], Подлесных С.В.[1], Колосова Е.А.[1], Щербаков Д.Н.[1], Петрова В.Д.[3], Джонстон С.А.[2], Лазарев А.Ф.[3], Оскорбин Н.М.[4], Шаповал А.И.[1,2], Рязанов М.А.[1]

[1]*Российско-американский противораковый центр, Алтайский государственный университет, Барнаул, Россия*
[2]*Центр инноваций в медицине, Институт биодизайна, Университет штата Аризона, Темпи, Аризона, США*
[3]*Алтайский филиал РОНЦ им. Н.Н. Блохина, Барнаул, Россия*
[4]*Факультет математики и информационных технологий, Алтайский государственный университет, Барнаул, Россия*

Аннотация. В настоящее время в качестве аналитической системы в различных медико-биологических исследованиях применяются биологические микрочипы, содержащие нуклеотиды, белки, пептиды, гликаны и другие биологические молекулы. Современные микрочипы активно модифицируются, увеличивается количество и плотность иммобилизированных молекул. Обработка больших массивов данных, полученных с помощью микрочипов, требует оптимизации алгоритмов их анализа. Данные получение на пептидных микрочипах имеют некоторые особенности и требует применения нестандартных методов статистического анализа. В настоящей работе представлены результаты анализа репертуара антител в сыворотках пациентов с диагнозом раком молочной железы, полученных с помощью микрочипов, содержащих 330 тысяч пептидов. Исследование методов уменьшения размерности, в частности, проекционных методов и методов отбора информативных признаков показало, что метод проекции на латентные структуры позволяет выявить эффективную размерность данных, уменьшить эффект переобучения модели и улучшить качество распознавания объектов. Точность результатов эксперимента оценена при помощи ROC-кривой, и наилучшее качество достигнуто с использованием трёх латентных структур без предварительной нормализации и с использованием всех пептидов.

Ключевые слова: *микрочипы, пептиды, нормализация, латентные переменные, кластеризация, ROC-кривая, метод проекции на латентные структуры.*

ВВЕДЕНИЕ

В современных медико-биологических исследованиях все большее распространение получает скрининговый подход. При этом актуальным направлением совершенствования методов лабораторной диагностики является создание микрочипов – многоэлементных матриц, миниатюрных аналитических устройств для одновременного анализа специфических взаимодействий биологических молекул [1–4].

[*]anisimow.d@gmail.com

Микрочип представляет собой твёрдый субстрат (стекло, бумага, пластик), поверхность которого содержит биологические молекулы (ДНК, белки, РНК, пептиды, моноклональные антитела, лиганды, рецепторы), клетки (млекопитающих или бактерий), специфически взаимодействующие с молекулами анализируемого образца с образованием комплексов [1, 5]. Виды иммобилизованных молекул определяют разнообразие микрочипов: белковые, клеточные, ДНК-микрочипы, пептидные и др. [6–9].

С помощью технологии микрочипов возможны изучение ДНК-белковых, белок-липидных, белок-лекарственных, белок-пептидных и белок-белковых взаимодействий, их биохимической активности и иммунного ответа [9–14]. На микрочипах проводится большое количество исследований, например, анализ патофизиологических процессов, поиск маркеров у больных в том числе при онкологических заболеваниях [15, 16]. Преимущества технологии микрочипов в сравнении с традиционными методами исследования – высокая скорость анализа больших данных, минимальное количество анализируемого биоматериала, единовременная детекция большого количества маркеров, низкая себестоимость исследования [4, 16]. Современные микрочипы активно модифицируются, увеличивается состав и количество (плотность) иммобилизированных молекул, повышается чувствительность и специфичность метода [17].

Нашей исследовательской группой разрабатывается метод ранней диагностики рака молочной железы (РМЖ). Метод основан на применении пептидных микрочипов со случайными аминокислотными последовательностями, которые используются для оценки взаимодействия антител с частичным или полным подобием эпитопов антигенов [10, 12, 14, 16]. Данный микрочип представляет собой кремниевую пластину (25 × 75 × 1 мм), которая содержит 24 идентичных микроматрицы по 330 034 пептида со случайными аминокислотными последовательностями на каждой. Пептиды синтезированы на подложке микрочипа методом фотолитографии. Аминокислотная последовательность и месторасположение каждого пептида на микроматрице известно. Каждая микроматрица имеет площадь 0.5 $см^2$, размер точки каждого пептида около 8 мкм в диаметре, расстояние между соседними пептидами приблизительно 1 нм [17].

В исследованиях с использованием микрочипов для выявления причин заболевания требуется проведение большого числа анализов и оценки множества межмолекулярных взаимодействий. Получаемые в экспериментальной работе с микрочипами данные хорошо воспроизводимы и представляют собой числовые данные. Обработка больших массивов данных, полученных на микрочипах, имеет некоторые сложности и требует применения нестандартных статистических методов анализа.

Следует отметить, что большинство методов анализа данных микрочипов разработаны для изучения РНК- или ДНК-микроматриц. Однако анализ данных белковых и пептидных микрочипов имеет свои уникальные особенности. Нуклеотиды (РНК или ДНК) обычно связываются только со специфически-комплементарными пробами, расположенными на микрочипе, в данном случае практически отсутствует неспецифическое взаимодействие с другими нуклеотидными последовательностями. Однако анализ взаимодействия антител (или других белков) с пептидами, расположенными на микрочипе, должен учитывать, что с одним пептидом могут взаимодействовать разные антитела. Также возможно, что другие белки, находящиеся в тестируемой сыворотке, могут связываться с пептидами и блокировать их взаимодействие со специфическими антителами. Это особенно проблематично, если сывороточные белки, блокирующие взаимодействие антител с пептидами, присутствуют в разных концентрациях у разных пациентов. Более того, антитело одной специфичности может связываться с разными пептидами на микроматрице, при этом

все связывания будут иметь разные аффинности, чего практически не наблюдается в случае нуклеотидных микроматриц.

Учитывая вышеизложенное, взаимодействие антител (или белков) с пептидами на микроматрице может быть подвержено влиянию многих факторов, и не понятно, могут ли существующие методы статистики быть использованы для адекватного анализа данных пептидных микроматриц. Основной задачей нашего исследования было определение наиболее подходящих методов статистического анализа, которые могут достоверно отличить сыворотки здоровых людей и больных с установленным диагнозом рака молочной железы. В дальнейшем эти методы, могут быть использованы при разработке алгоритмов для клинического установления диагноза.

Это актуализирует цель работы – применение метода проекции на латентные структуры для анализа больших данных на пептидных микрочипах.

МАТЕРИАЛЫ И МЕТОДЫ

Имеющиеся данные

Для исследования были сформированы две группы. Первая группа – 40 пациентов, женского пола (средний возраст 56.4 ± 12.2 года) с диагнозом рака молочной железы (РМЖ). Вторая группа (контроль) сформирована из 41 здорового донора женского пола (средний возраст 47.1 ± 8.5 года) без признаков РМЖ. Настоящее исследование было поддержано этическим комитетом Алтайского филиала РОНЦ им. Н.Н. Блохина (г. Барнаул). Участвующие в исследовании доноры и пациенты подписали письменное информированное согласие.

Биологическая часть эксперимента. Для анализа на микрочипах была использована капиллярная кровь. Забор крови из пальца производили в конические пробирки типа Микровет («Фирма Синтакон»), содержащие K_3ЭДТА. Пробирки с образцами центрифугировали. Оценку репертуара антител в сыворотках больных и здоровых доноров проводили с помощью микрочипов, содержащих 24 идентичных микроматрицы из 330 034 пептидов со случайными аминокислотными последовательностями.

Перед экспериментом микрочипы были специальным образом подготовлены. Для этого каждый микрочип на 60 мин отдельно помещали в дистиллированную воду, затем на 30 мин в фосфатно-солевой буфер (ФСБ, «Биолот»), инкубировали при малой скорости шейкера Biosan OS 20 («Biosan»). Далее микрочипы промывали полосканием в трёх свежих растворах ФСБТ (ФСБ + 0.25 % Твин 20, «Helicon») и дистиллированной воде. Высушенные центрифугированием, в течение 5 мин при 800 об./мин, микрочипы помещали в гибридизационную кассету («Arrait Corporation») с силиконовыми прокладками. В каждую лунку кассеты, соответствующей микроматрице микрочипа, добавляли по 150 мкл инкубационного раствора, содержащего ФСБТ и 3 % бычьего сывороточного альбумина (БСА, «Amresco»), инкубировали микрочипы в течение 18 ч при 4°C. Образцы исследуемой плазмы крови разводили (1:250) в инкубационном растворе и добавляли в лунки гибридизационной кассеты. Инкубировали в течение 60 мин на орбитальном шейкере при 250 об/мин. После инкубации, с помощью промывателя микропланшетов BioTek ELx50 («BioTek Instruments») микрочипы промывали тремя повторами свежего ФСБТ и промывочным раствором. Далее микрочипы помещали в четырёхлуночные планшеты, наполненные раствором «вторичных» антител с флюоресцентной меткой против иммуноглобулина человека класса G (IgG). Планшеты с микрочипами в растворе «вторичных» антител накрывали не пропускающей свет крышкой, инкубировали в течение 60 мин на малой скорости орбитального шейкера. После этого микрочипы, фиксированные в кюветах EasyDip, промывали полосканием в трёх свежих растворах ФСБТ, дистиллированной воде и в течение 5 мин высушивали центрифугированием (800 об./мин).

Компьютерный анализ данных. Высушенные микрочипы сканировали с использованием двухлазерного сканера высокого разрешения «InnoScan 900 AL» («Innopsys») при длине волн 632 нм и 535 нм. С использованием программного обеспечения Mapix (v. 7.3.1) было оцифровано 162 файла с интенсивностями флюоресценции каждого образца на микрочипах, содержащих 330 034 аминокислотных последовательностей (пептидов). Этот набор данных состоит из 82 файлов для контрольных доноров (КД) и 80 файлов для доноров с диагнозом рака молочной железы (для каждого донора было проведено по 2 технических повтора). Каждый файл содержит статистические данные по каждому пептиду, нанесённому на чип, например, среднее значение, медиана, стандартное отклонение люминесценции в области пептида (foreground) и вокруг пептида (background). К тому же данная статистика имеется в отдельности для двух длин волн – красной (635 нм) и зелёной (532 нм). Ввиду слишком большой размерности данных при малом количестве объектов было принято использовать одно значение для каждого пептида – медиана на длине волны 532 нм. Таким образом, для последующих экспериментов было отобрано 162 объекта по 330 034 исходных переменных, значения которых изменяются в диапазоне от 0 до 65535.

Предобработка

В качестве предварительной обработки использовались исследованные ранее методы [18, 19]. В частности, данные предварительно логарифмировались по основанию два, затем подвергались медианной нормализации для подавления отклонений фонового свечения различных чипов.

Используя предположение о том, что пептиды в технических повторах одного донора должны обладать одинаковой люминесценцией, делаем вывод о необходимости нормализации:

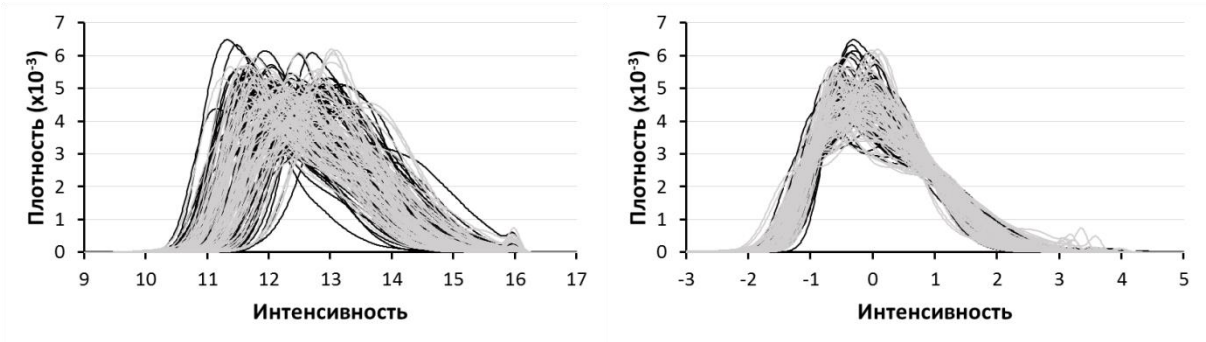
$$Il = \log 2(I + 1),\ In = Il - \mathrm{median}(Il),$$

где

I – интенсивность люминесценции пептидов, $I \in [0;65535]$;

Il – логарифм интенсивности, $Il \in [0;16]$;

In – нормализованный логарифм интенсивности, $In \in [-16;16]$.



а) б)

Рис. 1. Плотность распределения логарифма интенсивностей пептидов в каждом техническом повторе: а) – до нормализации, б) – после медианной нормализации. Серые линии – плотности распределений для доноров с диагнозом РМЖ, чёрные – для контрольных доноров. По оси *X* левая диаграмма – log2 интенсивности флюоресценции, правая диаграмма – условные единицы интенсивности флюоресценции после медианной нормализации.

Результат работы алгоритма нормализации показан на рисунке 1. Заметно, что после медианной нормализации (рис. 1,б) плотности распределений логарифма интенсивностей пептидов стали более однородными, чем до применения нормализации (рис. 1,а).

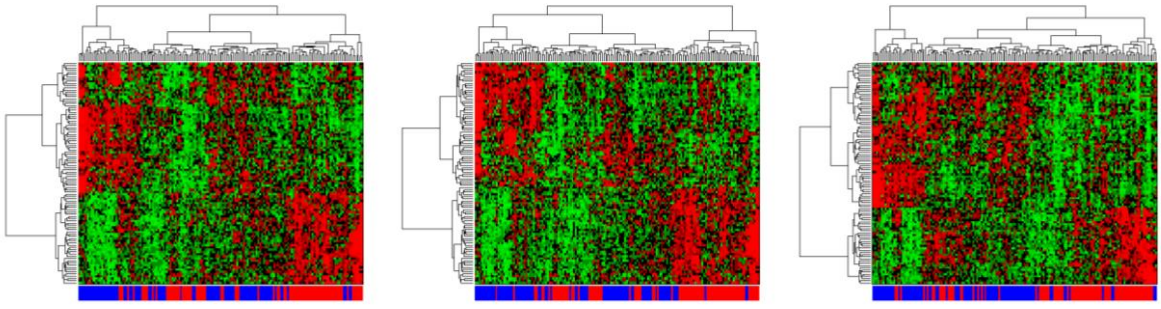
Уменьшение размерности

Существуют различные методы уменьшения размерности, например, проекционные (МГК, ПЛС) [20] и методы отбора информативных признаков (*t*-test, *U*-test) [21, 22]. Методы отбора информативных признаков позволяют выделить переменные, которые на основании определённого критерия наиболее важны (информативны) в задаче определения целевой переменной (класса). Проекционные методы позволяют выделять латентные переменные. В ряде случаев, в малом количестве таких переменных концентрируется наибольшая доля информации, содержащаяся в данных. Так же проекционные методы позволяют выявить эффективную размерность данных, удалить их шумовую составляющую и, как следствие, уменьшить переобучение моделей и улучшить качество распознавания объектов [20].

Для начала определимся с критерием информативности признаков. В данной работе рассматривалось три критерия: *t*-критерий Стьюдента, *U*-критерий Манна–Уитни и критерий, основанный на корреляции Пирсона. Для наглядности отбиралось небольшое количество пептидов, на основании каждого из критериев.

Принцип отбора информативных пептидов покажем на примере *t*-критерия Стьюдента. Для каждого пептида вычисляем *t*-статистику для равенства средних значений этого пептида в каждом классе доноров (КД/РМЖ). Чем больше значение *t*-статистики, тем более значимо различие средних между классами. Затем, сортируем все пептиды по полученному значению *t*-статистики в убывающем порядке и из начала данного списка выбираем необходимое количество пептидов.

Аналогичным образом поступаем с выбором информативных пептидов с использованием других критериев информативности, но вместо *t*-статистики нужно использовать значение, соответствующее выбранному критерию. Например, при использовании *U*-критерия для каждого пептида считаем значение *U* и затем выбираем определённое количество пептидов с наименьшим значением, так как чем меньше значение *U*, тем существеннее различие интенсивности пептидов между классами. Для критерия основанного на корреляции Пирсона, для каждого пептида считаем коэффициент корреляции между значением пептида и меткой класса (например, 0 для КД и 1 для РМЖ), затем выбираем необходимое количество пептидов с наибольшим рассчитанным значением, то есть наиболее коррелирующих с диагнозом.

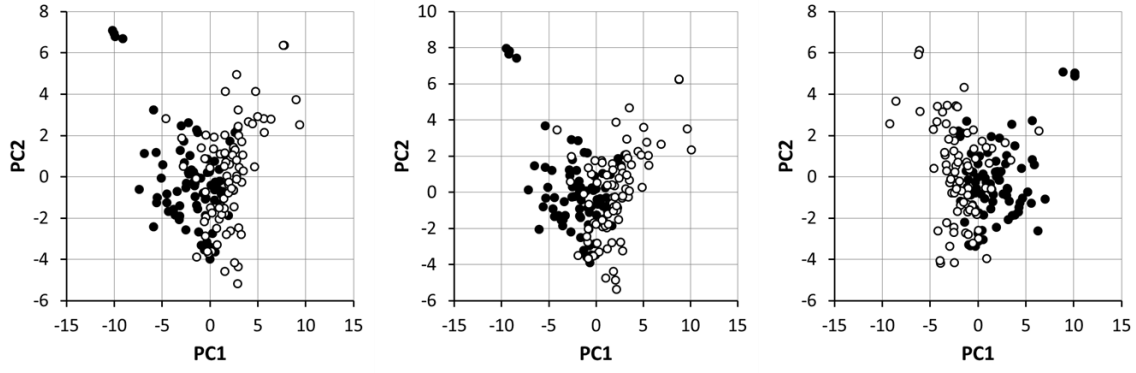


а) б) в)

Рис. 2. Иерархическая кластеризация и тепловая карта (heatmap) сывороток больных РМЖ и здоровых доноров. 100 Информативных пептидов, выбранных при помощи *t*-критерия Стьюдента (а), коэффициента корреляции Пирсона (б) или *U*-критерия Манна–Уитни (в) представлены в горизонтальных рядах. Образцы сывороток представлены в вертикальных колонках и обозначены синим (здоровые доноры) и красным (РМЖ пациенты) в нижней части рисунка. Красные точки на тепловой карте показывают высокое взаимодействие сыворотки с определенным пептидом, зеленые – показывают низкое взаимодействие. Кластеризация пептидов представлена на левой стороне рисунка. Кластеризация сывороток представлена в верхней части рисунка.

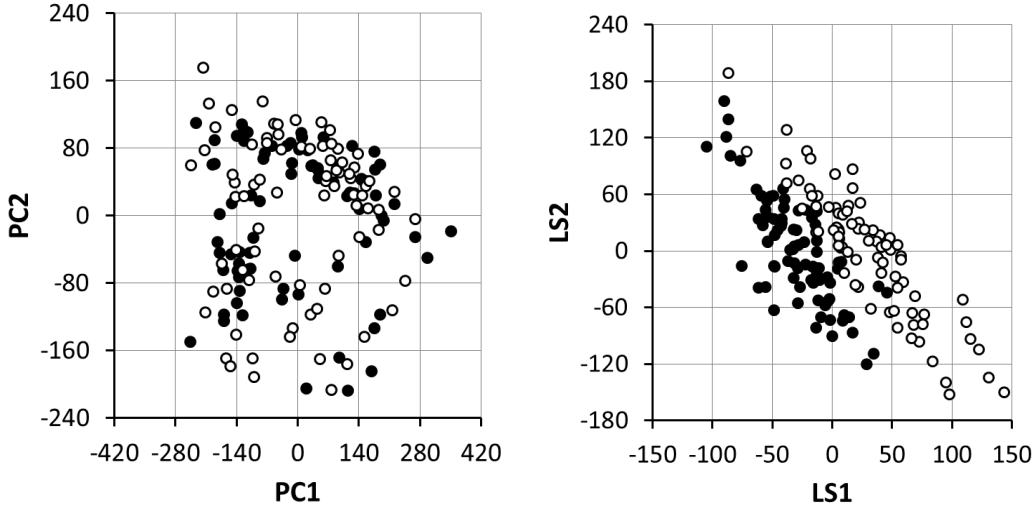
Качество разделения доноров на классы оценивалось визуально. Результат иерархической кластеризации доноров с использованием 100 пептидов, выбранных *t*-критерием, *U*-критерием и на основании корреляции показан на рисунках 2,а–в.

Визуализация данных методом главных компонент (МГК) для каждого способа отбора признаков представлена на рисунке 3. Наблюдаются качественно схожие результаты разделения групп объектов на классы РМЖ/КД. Воспользуемся более формальным признаком: *t*-критерий Стьюдента и критерий, основанный на корреляции, являются достаточно мощными в случае, когда данные в каждой группе распределены нормально [21], в то время как *U*-критерий Манна–Уитни не требует нормального распределения данных [22]. Но поскольку, на основании теста Лиллиефорса [23] с критическим уровнем значимости alpha = 0.05, в группе КД проверку на соответствие нормальному закону распределения проходят 60.37 % пептидов, а в группе РМЖ – 57.61 % пептидов, то для дальнейшего анализа остановимся именно на использовании *U*-критерия Манна–Уитни.



а) б) в)

Рис. 3. Визуализация результатов классификации образцов сыворотки больных РМЖ и здоровых доноров методом главных компонент (МГК) после отбора 100 информативных пептидов, с использованием *t*-критерия (а), коэффициента корреляции (б), *U*-критерия (в). Первые две главные компоненты (PC1 и PC2 от англ. principal components) представлены на осях *X* и *Y*, соответственно. Черные символы обозначают здоровых доноров, белые символы – пациенты с диагнозом РМЖ.



а) б)

Рис. 4. Результат уменьшения размерности данных с 330000 пептидов до двух латентных переменных: а) методом главных компонент; б) методом проекции на латентные структуры. Первые две главные компоненты (PC1 и PC2 от англ. principal components) представлены на осях *X* и *Y* левого графика, соответственно. Первые две латентные структуры (LS1 и LS2 от англ. latent structures) представлены на осях *X* и *Y* правого графика, соответственно.

После отбора информативных пептидов, большая их часть коррелирует друг с другом, что можно заметить на рисунке 2. То есть, данные имеют мультиколлинеарность от которой можно избавиться, используя проекционные методы уменьшения размерности, наиболее распространённым из которых является МГК. Но МГК не учитывает связь главных компонент с целевой переменной (рис. 4,а). Данную особенность учитывает метод проекции на латентные структуры (ПЛС) (рис. 4,б) [20].

Таким образом, для дальнейших расчётов будем использовать выделение информативных пептидов с помощью *U*-критерия Манна–Уитни с последующим уменьшением размерности методом ПЛС.

Протокол проведения эксперимента

Для оценки качества моделей использовалась перекрёстная проверка «один против всех». На *i*-й итерации выборка разбивалась на обучающую и тестовую так, что в тестовую выборку попадают все технические повторы *i*-го донора, а в обучающую остальные данные. Далее происходит выбор множества информативных пептидов на основании *U*-критерия. Затем, на выбранном множестве пептидов происходит настройка модели регрессии на латентные структуры, где в качестве независимых переменных использовались нормализованные интенсивности информативных пептидов *In*, а в качестве зависимой переменной выступала метка класса объектов.

Запишем модель регрессии на латентные структуры в следующем виде:

$$P_i = F_l(\mathbf{x}_i, \mathbf{b}),$$

где
l – количество латентных структур;
$\mathbf{b}$ – вектор параметров модели;
$\mathbf{x}_i$ – вектор переменных, характеризующих объект i;
P_i – целевая переменная для объекта i.

Оценку параметров модели $\mathbf{b}^*$ проведём в соответствии с подходом, описанным в [24]. Тогда задача обучения на выборке размером N может быть записана так:

$$\boldsymbol{b}^* = \arg\min_{\boldsymbol{b}} \frac{1}{N}\sum_{i=1}^{N}(F_l(\mathbf{x}_i, \mathbf{b}) - P_i)^2.$$

Теперь сведём задачу классификации доноров на классы КД/РМЖ к задаче регрессии. Положим класс объекта i равным $Y_i \in \{-1;1\}$ так, что $Y_i = 1$, если донор i имеет диагноз РМЖ, и $Y_i = -1$ если донор i является контрольным. Введём $P_i \in \mathrm{R}$ как меру подобия объекта i основному классу (РМЖ), и положим для объектов обучающей выборки $P_i = Y_i$. Затем, получив предсказанное моделью значение $\hat{P}_i$, можем назначить объекту i класс $\hat{Y}_i$ так, что $\hat{Y}_i = \mathrm{sgn}(\hat{P}_i - t)$, где $t \in \mathrm{R}$ – порог классификации, влияющий на баланс ошибок первого и второго рода.

После проведения полного цикла перекрёстной проверки, для объектов обучающей выборки имеем вектор истинных меток классов Y, предсказанных мер подобия $\hat{P}$ и предсказанных меток классов $\hat{Y}_i = \mathrm{sgn}(\hat{P}_i)$.

Определим меры качества классификатора:

$$E = \frac{1}{N}\sum_{i=1}^{N}\begin{cases}1, \hat{Y}_i \neq Y_i\\ 0, \hat{Y}_i = Y_i\end{cases};$$

$$Se = \frac{1}{N_{BC}}\sum_{i=1}^{N}\begin{cases}1, \hat{Y}_i = Y_i \text{ и } Y_i = 1\\ 0, \text{иначе}\end{cases};$$

$$Sp = \frac{1}{N_{ND}} \sum_{i=1}^{N} \begin{cases} 1, \hat{Y}_i = Y_i \text{ и } Y_i = 0 \\ 0, \text{иначе} \end{cases},$$

где

N – общее количество объектов выборки;

N_{BC} – количество объектов выборки, принадлежащих классу РМЖ;

N_{ND} – количество объектов выборки, принадлежащих классу КД;

E – совокупная ошибка, показывающая относительное количество объектов, неверно распознанных классификатором;

Se – чувствительность, то есть вероятность дать правильный ответ на объект основного класса (РМЖ);

Sp – специфичность, или вероятность дать правильный ответ для объекта дополнительного класса (КД).

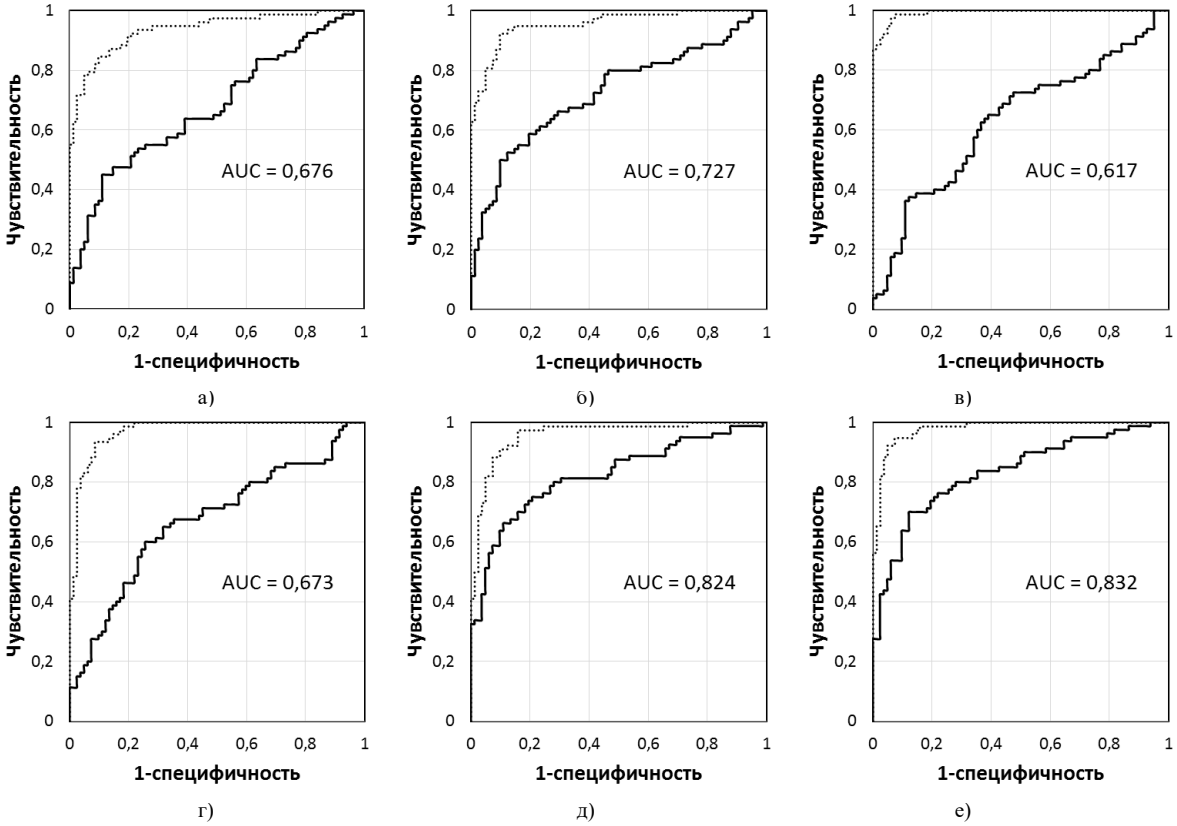


Рис. 5. ROC-кривые характеризующие классификатор основанный на ПЛС. Классификация производилась на данных после медианной нормализации и уменьшения размерности *U*-критерием до 100 переменных и двумя латентными структурами (а); 300 переменных и двумя латентными структурами (б); до 3000 переменных и тремя латентными структурами (в); до 30000 переменных и двумя латентными структурами (г); без предварительного уменьшения размерности и двумя латентными структурами (д); без нормализации, без предварительного уменьшения размерности и с тремя латентными структурами (е) Пунктирная линия – обучающая выборка, черная линия – тестовая выборка, в которой ошибкой считался каждый технический повтор. Цифрами обозначены AUC для тестовой выборки, большее значение указывает на лучшее качество классификатора.

Данные показатели являются точечными оценками качества классификатора. В ряде случаев интересно узнать, как классификатор реагирует на изменение баланса между ошибками I и II рода. Данная характеристика описывается кривой мощности критерия (ROC-кривой) и площадью под ней – ROC-AUC. Для расчёта ROC-кривой необходимо изменяя порог классификации t, назначать объектам класс и подсчитывать критерии качества Se и Sp. Затем получившееся множество точек наносим на график так, что на

Математическая биология и биоинформатика. 2017. Т. 12. № 2. doi: 10.17537/2017.12.435

оси абсцисс будут находиться значения 1 – *Sp*, а на оси ординат – соответствующие им значения *Se*.

Результаты эксперимента

Проведённые эксперименты показывают, что классические статистические методы уменьшения размерности негативно влияют на дальнейшее распознавание данных методом ПЛС. Это можно заметить на рисунках 5,а–д, где показаны ROC-кривые для классификаторов, построенных на данных с различным числом предварительно отобранных информативных пептидов. В этих экспериментах использование двух латентных переменных являлось оптимальным с точки зрения описанной выше совокупной ошибки (рис. 6). Однако, ROC-кривая, построенная по результатам классификации данных с медианной нормализацией и без уменьшения размерности (рис. 5,д), показывает сравнимую точность с классификатором, построенным на данных без предварительной нормализации (рис. 5,е), но в последнем случае, оптимальным являлось использование трёх латентных переменных (рис. 6).

Ярко выраженный минимум на графиках ошибок, изображённых на рисунке 6, может свидетельствовать о неустойчивости модели. Без нормализации минимум ошибки достигается при использовании трёх латентных структур, из-за чего можно сделать вывод о том, что алгоритм ПЛС способен учитывать отклонения в средних значениях уровня люминесценции пептидов, однако для этого требуется использование большего числа латентных переменных. Иными словами, нормализация уменьшает эффективную размерность данных.

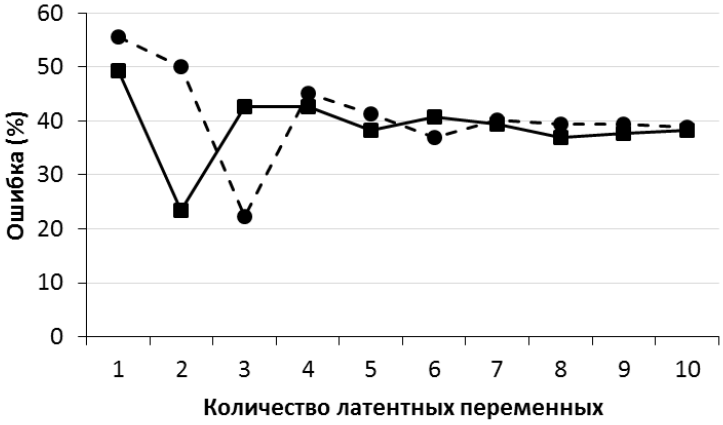


Рис. 6. Зависимость ошибки классификации от количества латентных переменных на всём множестве пептидов. (●) без нормализации, (■) с медианной нормализацией.

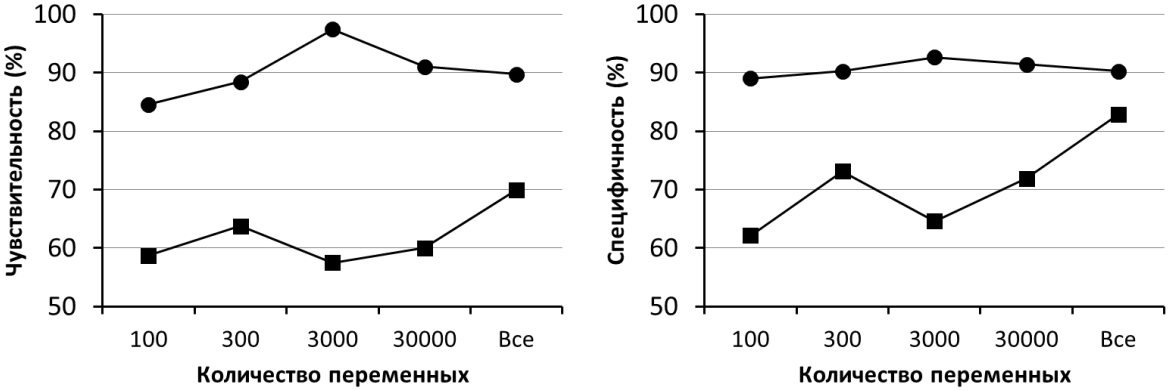


Рис. 7. Зависимость чувствительности и специфичности классификации методом ПЛС (проекция на латентные структуры) от количества информативных переменных. (●) Обучающая выборка, (■) Тестовая выборка. Классификация производилась с использованием данных после медианной нормализации, и уменьшением количества переменных с помощью *U*-критерия.

На рисунке 7 изображены показатели качества классификации, такие как чувствительность и специфичность, рассчитанные для различного количества информативных пептидов, отобранных с помощью *U*-критерия Манна–Уитни. Наилучшее качество по обоим параметрам достигается на всём множестве пептидов.

ЗАКЛЮЧЕНИЕ

В данной работе рассмотрена модель проекции на латентные структуры применительно к анализу данных пептидных микрочипов при ранней диагностике рака молочной железы и исследованы методы уменьшения размерности, в частности проекционные (МГК, ПЛС) и методы отбора информативных признаков (*t*-test, *U*-test). В работе показано, что проекционные методы позволяют выявить эффективную размерность данных, которая составляет две латентные переменные для предварительно нормализованных и три латентные переменные для данных без предварительной нормализации. Важной частью эксперимента являлось сокращение шумовой составляющей и, как следствие, уменьшение переобучения моделей, а также улучшение качества распознавания объектов.

Точность результатов эксперимента оценена при помощи ROC-кривой и наилучшее качество достигнуто с использованием трёх латентных структур без предварительной нормализации и уменьшения размерности. При этом ошибка классификации была на уровне 21 %, чувствительность 70 %, а специфичность более 87 %.

Рассматриваемый в данной работе подход будет использован для создания автоматизированного рабочего места лаборанта, которое позволит более эффективно проводить предварительный анализ результатов с использованием технологии пептидных микрочипов.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 17-04-00321 и в рамках выполнения государственного задания Минобрнауки России №6.3892.2017/4.6.

СПИСОК ЛИТЕРАТУРЫ

1. Осипова Т.В., Рябых Т.П., Барышников А.Ю. Диагностические микрочипы: применение в онкологии. *Российский биотерапевтический журнал*. 2006. Т. 5. № 3. С. 72–81.
2. Никитин Е.А., Судариков А.Б., Баранова А.В. Микрочипы: новый этап в онкогематологии. *Онкогематология*. 2008. № 1–2. С. 6–12.
3. Кожевникова О.С., Мартыщенко М.К., Генаев М.К., Корболина Е.Е., Муралева Н.А., Колосова Н.Г., Орлов Ю.Л. RatDNA: база данных микрочиповых исследований на крысах для генов, ассоциированных с заболеваниями старения. *Вавилов. журн. генет. и селекции*. 2012. Т. 16. № 4/1. Р. 756–765.
4. Осипова Т.В., Рябых Т.П., Дементьева Е.И., Дарий Е.Л., Рубина А.Ю., Заседателев А.С. Белковые микрочипы для диагностики злокачественных новообразований. Разработка биочипа на простата-специфический антиген. *Российский биотерапевтический журнал*. 2003. Т. 2. № 3. С. 24–30.
5. Наседкина Т.В. Использование биологических микрочипов в онкогематологии. *Онкогематология*. 2006. № 1–2. С. 25–37.
6. Китаева Н.В., Фриго Н.В., Волков И.А., Лихарёва В.В. Биомикрочипы и возможность их применения в дерматовенерологии. *Вестник дерматологии и венерологии*. 2009. № 6. С. 33–45.
7. Hall J. The microarray revolution: how one chip is changing the face of science. *Harvard Science Rev*. 2002. P. 82–85.

8. Jain K.K. The role of protein chip technology in molecular diagnostics. *IVD Technology*. 2002. V. 8. P. 49–56.
9. Kijanka G., Murphy D. Protein arrays as tools for serum autoantibody marker discovery in cancer. *J. Proteomics.* V. 72. № 6. P. 936–944.
10. Podlesnykh S.V., Kolosova E.A., Shcherbakov D.N., Shaidurov A.A., Anisimov D.S., Ryazanov M.A., Johnston S.A., Shoikhet Ya.N., Petrova V.D., Lazarev A.F., Chapoval A.I. Interaction of serum antibodies from breast cancer patients with synthetic peptides. *Bulletin of Experimental Biology and Medicine*. 2016. V. 161. No. 6. P. 816–820. doi: 10.1007/s10517-016-3519-7
11. Stafford P., Brun M. Three methods for optimization of cross-laboratory and cross-platform microarray expression data. Nucleic Acids Res. 2007. V. 35. № 10. P. 1-16.
12. Rubina A.Y., Dementieva E.I., Stomakhin A.A., Darii E.L., Pankov S.V., Barsky V.E., Ivanov S.M., Konovalova E.V., Mirzabekov A.D. Hydrogel – based protein microchips: manufacturing, properties, and applications. *BioTechniques.* 2003. V. 34. P. 1008-1022.
13. Hardiman G. Microarray platforms -- comparisons and contrasts. *Pharmacogenomics.* 2004. V. 5. № 5. P. 487-502.
14. Lacombe J., Mangé A., Solassol J. Use of autoantibodies to detect the onset of breast cancer. *Journal of Immunology Research*. 2014. P. 8.
15. Blohm D.H., Guiseppi-Elie A. New developments in microarray technology. *Curr. Opin. Biotechnol*. 2001. V. 12. P. 41–47.
16. Шаповал А.И., Легутки Д.Б., Стаффорд Ф., Требухов А.В., Джонстон С., Шойхет Я.Н., Лазарев А.Ф. Иммуносигнатура (immunosignature) – пептидные микроэррей для диагностики рака и других заболеваний. *Российский онкологический журнал*. 2014. № 4. С. 6–11.
17. Legutki J.B., Zhao Z.G., Greving M., Woodbury N., Johnston S.A., Stafford P. Scalable high-density peptide arrays for comprehensive health monitoring. *Nature Communications*. 2014. V. 5. P. 4785.
18. Анисимов Д.С., Рязанов М.А., Шаповал А.И. Подход к обработке многомерных данных пептидных микрочипов. *Известия АлтГУ.* 2015. Т. 1/2. № 85. С. 77–80. doi: 10.14258/izvasu(2015)1.2-13
19. Cretich M., Chiari M. *Peptide Microarrays: Methods and Protocols.* Humana Press. 2009. doi:10.1007/978-1-60327-394-7
20. Эсбенсен К. *Анализ многомерных данных. Избранные главы*. Барнаул: Изд-во Алт. ун-та, 2003. 157 с.
21. Student. The probable error of a mean. *Biometrika*. 1908. V. 6. № 1. P. 1–25.
22. Mann H.B., Whitney D.R. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*. 1947. № 18. P. 50–60.
23. Lilliefors H. On the Kolmogorov–Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*. 1967. V. 62. № 318. P. 399–402.
24. Максимов А.В., Оскорбин Н.М. *Многопользовательские информационные системы: основы теории и методы исследования*. Барнаул: Изд-во Алт. ун-та, 2013. 264 с.

Рукопись поступила в редакцию 04.07.2017.
Дата опубликования 29.11.2017.