УДК 57.087.1, 51.76

Т. Р. Жангиров, А. С. Перков, А. А. Лисс
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В. И. Ульянова (Ленина)

Н. Ю. Григорьева
Санкт-Петербургский государственный университет

Л. В. Чистякова
РЦ «Культивирование микроорганизмов» НП Санкт-Петербургского государственного университета

Применение линейного дискриминантного анализа для классификации цианобактерий по спектрам собственной флуоресценции

На основе применения специальных методов математической статистики разработана методика классификации различных родов цианобактерий по спектрам собственной флуоресценции отдельных клеток in vivo. Проведен анализ нескольких сотен спектров, полученных средствами конфокальной лазерной сканирующей микроскопии для 20 различных штаммов цианобактерий, и разработана определенная процедура обработки исходных данных, а также порядок определения параметров, отражающих как общую форму спектров, так и соотношение отдельных пиков. Статистическими методами определен ряд ключевых параметров, достаточных для проведения процедуры классификации. Для решения задачи классификации применен стандартный многомерный дискриминантный анализ. Результаты работы предложенного алгоритма классификации продемонстрированы на примере дифференциации трех штаммов, принадлежащих к двум родам цианобактерий.

Цианобактерии, дискриминантный анализ, статистические методы в биологии, спектры флуоресценции

В экологических исследованиях для определения уровня загрязнения водоемов биологическими методами используют преимущественно данные о видовом разнообразии организмов, формирующих сообщества, и их количественном соотношении. Одной из основных групп, формирующих фитопланктон и способных вызывать опасные токсичные «цветения» открытых водоемов, являются цианобактерии. В связи с этим актуальной представляется разработка новой методики, ориентированной на специфические особенности фотосинтетического аппарата цианобактерий и позволяющей оперативно оценивать видовой состав организмов в сообществе. Традиционно, видовое разнообразие определяется с помощью прямого микроскопирования проб, а биомасса фитопланктона – методом экстракции хлорофилла. Эти методы достаточно трудозатратны, что затрудняет их применение в задачах непрерывного экологического мониторинга отдельных водоемов из-за огромных объемов данных, подлежащих обработке.

Альтернативными методами дифференциации видов фитопланктона в натурных пробах являются химический и спектральный анализ. Одним из основных химических методов для таксономического анализа является метод жидкостной хроматографии [1], [2], однако он также не может обеспечить достаточно высокую скорость обработки информации. Напротив, методы спектрофотометрии могут обеспечить достаточно быстрый и относительно дешевый доступ к необходимой информации, и, кроме того, получаемые этим способом данные удобны для дальнейшей математической обработки.

Спектрофотометрические методы включают в себя исследования спектров поглощения, собственной флуоресценции и возбуждения флуо-

ресценции фотосинтетического аппарата микроводорослей. Большинство опубликованных по данной тематике работ базируется на анализе спектров поглощения отдельных видов фитопланктона [3], [4]. В формировании спектров поглощения участвуют все фотопигменты, входящие в состав фотосинтетического аппарата, причем форма спектра достаточно сильно зависит от концентрации отдельных элементов. Проблема заключается в том, что данным способом легко дифференцировать большие классы микроводорослей, но практически невозможно провести дифференциацию внутри одного класса по видам и штаммам, так как спектры поглощения даже для отдельных видов внутри одного класса слабо отличаются, что и отмечается в указанных статьях.

В литературе представлен также ряд исследований, основанных на анализе спектров собственной флуоресценции различных видов и классов фитопланктона [5]–[8]. Однако все предыдущие попытки проведения таксономического анализа различных видов микроводорослей по оптическим спектрам основывались на изучении интегральных спектров культуры в целом. Очевидно, что соотношение концентраций вспомогательных пигментов, таких, как каротиноиды и фикобилипротеины, а следовательно, и форма спектров могут достаточно быстро меняться в связи со старением отдельных клеток культуры. Кроме того, в интегральные спектры флуоресценции дают вклад и культуральная среда, и продукты жизнедеятельности цианобактериальной культуры, поэтому спектры культур одного штамма, выращенных при различных условиях, могут достаточно сильно отличаться друг от друга. В связи с этим на основании интегральных спектров исследователям удавалось разделить только те крупные группы фитопланктона, которые сильно различались по исходному набору основных фотопигментов.

Таким образом, в основе известных к настоящему времени спектроскопических методов лежит дифференциация таксонов на основе разного количественного и качественного состава фотоактивных пигментов в фотосинтетическом аппарате различных видов фитопланктона. Так, например, оптические свойства цианобактерий определяют 3 основных вида фотопигментов – это хлорофилл *a*, каротиноиды и фикобилипротеины.

В описываемой работе для дифференциации различных родов и штаммов цианобактерий использовались спектры собственной флуоресценции отдельных живых клеток, полученные средствами конфокальной микроспектроскопии, а не традиционными методами флуоресцентной спектроскопии для всей культуры в целом. Для анализа отбирались клетки, находящиеся в хорошем физиологическом состоянии, спектры которых предположительно отражают работу фотосинтетического аппарата и его молекулярное строение для данного рода/штамма наилучшим образом. Кроме того, при анализе использовались наборы спектров собственной флуоресценции одной клетки, полученные при возбуждении разными длинами волн видимого диапазона. В этом и состоит принципиальное отличие данных исследований от всех предыдущих. Предполагалось, что использование серии спектров, полученных с единичной клетки, для характеристики каждого штамма позволит выявить тонкие различия в строении светособирающих комплексов фотосинтетической системы между отдельными родами цианобактерий и даст более полную информацию для проведения статистического анализа.

Различные статистические методы для задач классификации применяются во многих областях научных исследований: в экономике, социологии, медицине [9], криминалистике, биологии. Наиболее корректным методом для решения поставленной задачи оказался линейный дискриминантный анализ [10]. Для извлечения из исходных данных параметров, пригодных для применения в дискриминантном анализе, был разработан алгоритм и реализована программа с использованием математического пакета MATLAB. Дальнейший анализ проводился в специализированном математическом пакете STATISTICA.

Таким образом, цель описываемой работы состояла в разработке и применении метода оперативного выявления и первичной дифференциации различных родов и штаммов цианобактерий на основе спектров собственной флуоресценции отдельных клеток in vivo. В качестве метода классификации был выбран линейный дискриминантный анализ. Для демонстрации эффективности разработанной методики были отобраны данные по трем штаммам цианобактерий, принадлежащих двум родам Leptolyngbya и Geitlerinema, и исследовано качество межродового и внутриродового разделения.

Постановка задачи. Объектом статистического анализа являются экспериментальные данные, полученные на конфокальном лазерном сканирующем микроскопе Leica TCS-SP5. Данные представляют собой серии из восьми спектров

собственной флуоресценции, снятые с отдельных клеток цианобактерий in vivo с помощью стандартной операции лямбда-сканирования программного обеспечения «Leica Confocal Software». Спектры снимались при различных длинах волн возбуждающего излучения, соответствующих лазерным линиям конфокального микроскопа: 405, 458, 476, 488, 496, 514, 543 и 633 нм. На рис. 1 представлена характерная серия спектров собственной флуоресценции клеток цианобактерий штамма Leptolyngbya CALU 1715. Каждый спектр нормирован на единицу и сдвинут по горизонтальной оси на 90 нм относительно соседнего для удобства рассмотрения серий спектров. Цифры над кривыми указывают длину волны возбуждающей лазерной линии. Штриховые линии отмечают ориентировочные положения максимумов флуоресценции пигмент-белковых комплексов (656, 682 и 715 нм).

С формальной точки зрения каждый спектр представляет собой массив чисел, интенсивностей флуоресценции, в диапазоне 560…785 нм с шагом 6 нм. В качестве примера в данной статье рассматриваются данные для трех штаммов цианобактерий: 1713, 1715, 1718 из коллекции CALU РЦ «Культивирование микроорганизмов» НП СПбГУ, которые относятся к родам Leptolyngbya (1713, 1715) и Geitlerinema (1718), для того чтобы можно было подробно исследовать качество как межродового, так и внутриродового разделения. Для достижения статистической достоверности выборки для каждого штамма было получено 25–30 наблюдений.

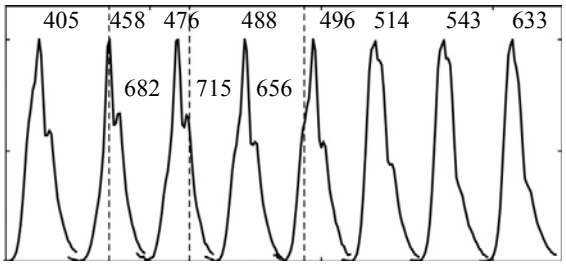


Рис. 1

На вход выбранного метода классификации подаются наблюдения с предварительно извлеченным набором параметров (порядка 80), для которых заранее известно, к какому классу они принадлежат. В связи с биологическим характером задачи параметры не являются независимыми и между ними возможна корреляция. В результате проведения классификации различные наблюдения должны быть приписаны к одному из изначально заданных классов, т. е. родов цианобактерий. Для данной задачи требования, предъявляемые к методу классификации, следующие: устойчивость результатов классификации, работа с априорной информацией, использование статистических оценок, простота реализации и наглядность представления результатов.

Процедура обработки исходных данных. Амплитуда сигнала и соотношение сигнал/шум в исходных экспериментальных данных, а также возможная отстройка выходного фильтра по частоте в значительной степени зависят от характеристик конкретного конфокального микроскопа. Однако поскольку в данной задаче основным параметром является форма спектров, то все прибор-зависимые флуктуации можно исключить на начальной стадии исследования. Поэтому для корректности дальнейшего статистического анализа необходимо провести предварительную обработку исходных данных.

Процедура предварительной обработки состоит из шести этапов:

1. Интерполяция спектров с шагом 1 нм. Это позволяет повысить точность извлекаемых параметров. Так как спектры флуоресценции (рис. 1) не имеют сильных осцилляций и достаточно гладкие, то используется интерполяция кубическими сплайнами.

2. Нормировка на максимальную интенсивность абсолютных значений спектров для устранения возможной флуктуации мощности возбуждающих лазеров.

3. Обрезание краев спектров, не несущих полезной информации, – исключение сильных шумовых флуктуаций в областях малой интенсивности излучения. Обрезание производится по значениям первой производной (слева – больше нуля, справа – меньше нуля).

4. Для исключения возможной ошибки определения полосы частот выходного спектра, возникающей при недостаточно прецизионной настройке конфокального микроскопа, производится центрирование исходного спектра. Эта процедура производится по спектру, соответствующему длине волны возбуждения 458 нм, и центрирование (т. е. сдвижка в ноль по оси абсцисс) идет по пику флуоресценции, соответствующему излучению хлорофилла *a* (~682 нм). Спектры для других лазерных линий в рамках данной серии также сдвигаются на соответствующие значения.

5. В ходе исследований было обнаружено, что значения извлекаемых параметров могут сильно зависеть от количества экспериментальных точек и различия значений интенсивности на границах интервала. Для устранения данной ошибки была введена дополнительная экстраполяция спектров на основе кубических сплайнов. На выходе этой процедуры все спектры лежат в диапазоне от –300 до 300 и на краях имеют нулевые значения.

6. На последнем этапе производится двухступенчатое сглаживание спектров для устранения малых колебаний, обусловленных тепловым шумом при снятии спектральных характеристик во время эксперимента. Используется метод скользящего среднего. На первой стадии применяется прямоугольное окно шириной 11 нм, а на втором – окно Ханна шириной 15 нм. Данные размеры оконных функций были выбраны как оптимальные, так как использование больших размеров сильно искажает форму спектров, что приводит к потере информации, а использование меньших размеров окна не дает удовлетворительного результата.

Использование данной процедуры обработки исходных данных позволяет устранить возможные аппаратные ошибки. Для реализации предварительной обработки исходных данных была разработана соответствующая программа в среде MATLAB. На выходе каждый спектр в серии из восьми штук представляет собой массив чисел размерностью 600 × 2 $\left\{x^{\lambda}, y^{\lambda}\right\}$, где λ – длина волны возбуждающего лазера (405, 458, 476, 488, 496, 514, 543 и 633 нм).

Описание параметров для статистического анализа. Поскольку для решения задачи классификации цианобактерий по спектрам собственной флуоресценции основным дискриминирующим параметром является форма спектров, т. е. общая площадь под кривой и соотношение пиков в частотных областях излучения основных пигментов фотосинтетического аппарата, то необходимо определить ряд параметров, наиболее полно отражающих эти характеристики. В описываемой работе в качестве таких параметров были выбраны: отношения максимумов флуоресценции различных пигментов; относительная интенсивность флуоресценции в различных спектральных областях; параметры асимметрии и эксцесса для каждого спектра. Таким образом, при использовании данных для всех восьми лазерных линий возбуждения для каждого экспериментального наблюдения получаем набор из 88 параметров, на основе которых в дальнейшем производится классификационный анализ.

Отношение интенсивностей флуоресценции различных пигментов. Этот параметр отражает тот факт, что разные штаммы цианобактерий имеют различный набор фотоактивных пигмент-белковых комплексов и различную эффективность их связывания между собой. Следовательно, соотношение плеч и пиков на спектрах флуоресценции, соответствующих определенным пигмент-белковым комплексам, будет различно. Для расчета данного параметра каждый спектр разбивается на 4 диапазона ([–149, –69], [–68, –8], [–7, 31], [32, 101]), и в каждом диапазоне ищется максимальное значение. Если значение попадает на край интервала, то берется среднее значение на данном интервале. Далее находятся соответствующие отношения интенсивностей по формуле

$$R_N^{\lambda} = \frac{M_N^{\lambda}}{M_3^{\lambda}},$$

где M_N^{λ} – максимальная интенсивность флуоресценции на промежутке $N = [1, –4]$ – номер области.

Отношения находятся относительно максимальной интенсивности флуоресценции в области 3, которая соответствует излучению хлорофилла *a*. Таким образом, для каждого наблюдения получается 24 параметра.

Относительная интенсивность флуоресценции отдельных пигментов. Этот параметр отражает процентное содержание интенсивности флуоресценции отдельных пигментов в общей флуоресценции фотосинтетического аппарата. Для вычисления данного параметра на спектре выделяется 5 областей ([–106, –95], [–41, –36], [–27, –23], [–3, 4], [33, 42]), на которых рассчитывается среднее значение интенсивности μ_k^{λ}, а затем процентное содержание P_k^{λ}:

$$\mu_k^{\lambda} = \frac{\sum_i y_{ki}^{\lambda}}{L_k}, \; P_k^{\lambda} = \frac{\mu_k^{\lambda}}{\sum_k \mu_k^{\lambda}},$$

где L_k – количество точек в области $k = [1, –5]$ – номер области.

В отличие от предыдущего, данный параметр рассматривает усредненные значения интенсив-

ности флуоресценции в информационно-полезных областях, что уменьшает влияние шумовых флуктуаций спектра. Кроме того, рассматривая процентный вклад каждой области в общую интенсивность флуоресценции, тем самым опосредованно учитывается и общий характер кривой – ее острота и асимметричность. Для каждого наблюдения получается 40 параметров.

Асимметрия и эксцесс. Эти параметры позволяют оценить общую форму спектра, его асимметричность и остроту. Понятно, что данные параметры не дают возможности разделить штаммы внутри одного рода, однако они позволяют разделить выборку на большие группы по основным родовым признакам, а в качестве дополнительных параметров к двум предыдущим группам улучшают результат классификации. Параметр асимметрии позволяет оценить наличие левого или правого пика:

$$A^{\lambda} = \sum_{i=-300}^{300}\left[\left(x_i^{\lambda} - \sum_{j=-300}^{300}\left(x_j^{\lambda} y_j^{\lambda}\right)\right)^3 y_i^{\lambda}\right],$$

а параметр эксцесса позволяет оценить ширину спектра:

$$E^{\lambda} = \sum_{i=-300}^{300}\left[\left(x_i^{\lambda} - \sum_{j=-300}^{300}\left(x_j^{\lambda} y_j^{\lambda}\right)\right)^4 y_i^{\lambda}\right] - 3.$$

Данные параметры требуют дополнительной нормировки исходных спектров таким образом, чтобы выполнялись следующие условия: $y_i^{\lambda} \geq 0$, $\sum_i y_i^{\lambda} = 1$ при $i = -300...300$.

Для каждого наблюдения получается еще 16 параметров.

Выбор оптимальной группы параметров. Поскольку суммарное количество параметров по всем трем группам для каждого наблюдения равно 80, то необходимо определить, какие из них являются критическими, а какие малозначащими, коррелирующими и, возможно, ухудшающими процедуру классификации. Для этого в математическом пакете STATISTICA был проведен анализ статистических характеристик для каждого параметра, при этом оценивались его математическое ожидание и дисперсия для данной выборки наблюдений. В результате были исключены те параметры, у которых дисперсия по всем классам близка к нулю, а также те, у которых средние значения и дисперсия по всем классам совпадают со средним значением и дисперсией отдельного класса. Эти параметры не несут никакой полезной информации в плане проведения дальнейшей классификации. Такими параметрами оказались отношения интенсивностей и относительные интенсивности для соответствующих первых диапазонов. Это объясняется тем, что для выбранных в качестве примера штаммов пигменты, имеющие флуоресценцию в данной области, отсутствуют.

Далее оставшиеся 64 параметра были проверены на наличие взаимосвязей (сильных корреляций) между ними, что также может затруднить процесс классификации. Для этого была проведена иерархическая кластеризация по взвешенному центроидному методу. По результатам оценки евклидовых расстояний между параметрами были выявлены наборы параметров, близких друг к другу, которые могут быть заменены одним параметром из данной группы. Например, параметры, соответствующие четвертой области для лазерных линий 514 и 633 нм, имеют практически нулевое расстояние, поэтому один из таких параметров может быть исключен из рассмотрения. Таким образом, для таких групп параметров, как «отношение интенсивностей» и «относительная интенсивность», были исключены параметры, соответствующие линиям лазеров 476, 496, 514 и 633 нм, так как оказалось, что данные параметры несут информацию, практически эквивалентную линиям 488, 405 и 543 нм соответственно.

В результате исследования было установлено, что исходные экспериментальные данные избыточны. В измерениях присутствуют спектры, соответствующие лазерным линиям возбуждения, которые не несут дополнительной информации. Этот факт является ключевым при последующем создании автоматизированного программно-аппаратного комплекса для экологического мониторинга, так как число возбуждающих линий может быть уменьшено вдвое без потери эффективности классификации.

Таким образом, общее количество параметров для каждого наблюдения было уменьшено с 80 до 36.

Для извлечения из исходных спектров необходимого набора параметров была разработана соответствующая компьютерная программа с использованием математического пакета MATLAB.

Выбор метода классификации. С математической точки зрения задачу определения родовой принадлежности цианобактерий можно решать двумя различными группами методов. Первая группа – это методы кластерного анализа, которые позволяют найти естественное разделение в пространстве параметров. К ним относятся такие методы, как метод *k*-средних [11] и иерархическая кластеризация [12]. Вторая группа – методы классификации, которые позволяют найти закономерность между признаками и целевым классом. К ним относятся деревья классификации [13], наивный байесовский классификатор [14] и линейный дискриминантный анализ [10].

При использовании методов первой группы алгоритм определения родовой принадлежности сводится к предварительному построению кластеров на основе известных данных, а для нового наблюдения сводится к нахождению ближайшего к нему кластера. Очевидные минусы такого подхода – необходимость хранить информацию обо всех наблюдениях, при этом не оцениваются статистические характеристики, наблюдается плохая устойчивость к статистическим погрешностям и резким скачкам в исходных данных. Использование методов классификации позволяет избавиться от основных недостатков методов кластеризации, так как в данном случае модель для определения целевого класса строится на основе статистических характеристик и объем хранимой информации зависит только от количества признаков и не зависит от количества наблюдений. Для алгоритмов классификации, как и для алгоритмов кластеризации, модель строится на основе уже известных наблюдений. Основным недостатком алгоритмов классификации является то, что качество модели зависит от количества наблюдений, т. е. чем больше известных наблюдений, тем точнее будет оценка статистических характеристик и тем правильнее будет проводиться дифференциация групп. Кроме того, на эффективность работы модели существенно влияют корреляции в исходных данных.

Сравнение различных методов и выбор оптимального метода для решения поставленной задачи проводились по нескольким критериям. Результаты представлены в табл. 1. По результатам сравнения был выбран линейный дискриминантный анализ, так как он позволяет учитывать статистические характеристики исходных данных, результаты его классификации стабильны, в отличие от байесовского классификатора, он допускает наличие слабой корреляции, а также позволяет графически отобразить результаты классификации посредством канонического дискриминантного анализа Фишера.

Линейный дискриминантный анализ (ЛДА) – это алгоритм классификации, основывающийся на характеристиках распределения групп объектов, таких, как математическое ожидание и дисперсия (матрица ковариации). Основная идея заключается в построении линейной функции (полинома первой степени), зависящей от параметров для каждого класса. Преимуществом метода является то, что функции можно построить, не имея эталонных объектов, а зная только статистические характеристики каждого класса. Для классификации объекта необходимо рассчитать значение каждой разделяющей функции, при этом функция, дающая максимальное значение, определяет класс объекта. Известно, что наилучший результат классификации достигается, если параметры классов распределены по нормальному закону и имеют одинаковые матрицы ковариации.

С ЛДА тесно связан линейный дискриминантный анализ Фишера [10], который позволяет пони-

Таблица 1

Метод	Использование статистических характеристик	Необходимость отсутствия корреляций	Работа с категориальными переменными	Возможность графического представления	Минимизация хранимой информации
k-средних	Нет	Нет	Нет	Нет	Нет
Иерархическая кластеризация	Нет	Нет	Да	Да	Нет
Деревья классификации	Нет	Нет	Да	Да	Да
k-ближайших соседей	Нет	Нет	Нет	Нет	Нет
Наивный байесовский классификатор	Да	Да	Нет	Нет	Да
ЛДА	Да	Частично	Нет	Да	Да

зить размерность пространства параметров, построив новый ортогональный базис, а также графически оценить качество классификации ЛДА. Базис строится так, чтобы максимизировать расстояние между классами и минимизировать расстояния между наблюдениями внутри одного класса.

Результаты применения линейного дискриминантного анализа. Анализ проводился в математическом пакете STATISTICA на основе экспериментальных данных, соответствующих спектрам собственной флуоресценции при возбуждении 405, 458, 488 и 543 нм лазерными линиями. Классификация проводилась для 78 наблюдений из трех классов по 8, 12 и 20 параметрам. В данную выборку для всех трех штаммов были намеренно включены наблюдения (т. е. серии спектров флуоресценции), полученные с клеток в измененном физиологическом состоянии. Это было сделано для того, чтобы показать устойчивость выбранного метода классификации к возможным отклонениям в экспериментальных данных. В качестве метода классификации был выбран линейный дискриминантный анализ. Критерием качества классификации в данном случае является процент правильно определенных штаммов для всех рассматриваемых наблюдений, а также только для наблюдений, принадлежащих одному роду. Графически качество дифференциации штаммов можно продемонстрировать, используя канонический дискриминантный анализ Фишера, который позволяет спроецировать наблюдения из исходного пространства параметров в пространство меньшего размера, максимизировав расстояния между классами и минимизировав расстояния между наблюдениями внутри одного класса.

На рис. 2 представлены результаты дискриминантного анализа для групп параметров «асимметрия и эксцесс» (рис. 2, *а*), «отношение интенсивностей» (рис. 2, *б*), «относительные интенсивности» (рис. 2, *в*), по отдельности, а также при совместном применении групп «относительные интенсивности» и «асимметрия и эксцесс» (рис. 2, *г*).

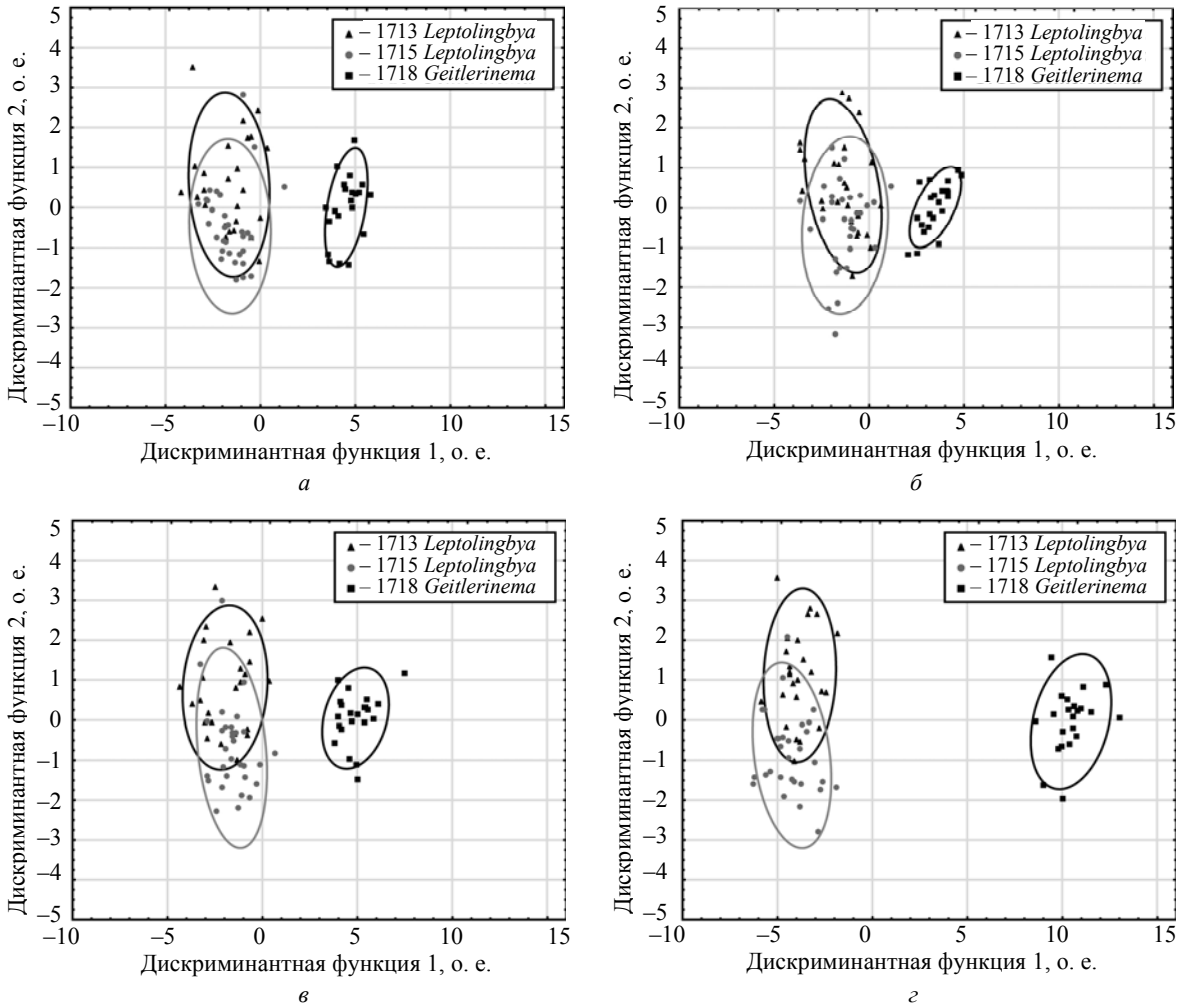


Рис. 2

Таблица 2

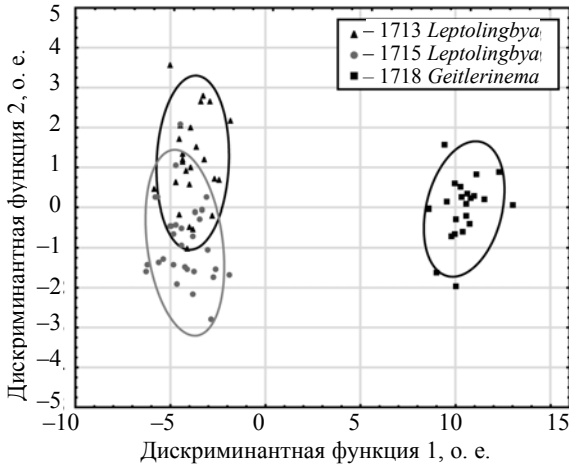
№	Наименование группы параметров	1713	1715	1718	Всего
	Для ограниченного числа лазерных линий (405, 458, 488, 543 нм)				
1	Асимметрия и эксцесс (8)	64.00 % (53.33)	70.97 % (75.86)	100 % (100)	76.92 % (77.78)
2	Отношения интенсивностей (8)	64.00 % (73.33)	64.52 % (68.97)	100 % (100)	74.36 % (79.37)
3	Относительные интенсивности (12)	76.00 % (80.00)	83.87 % (86.21)	100 % (100)	85.90 % (88.89)
4	Отношения интенсивностей + асимметрия и эксцесс (16)	64.00 % (66.67)	77.42 % (79.31)	100 % (100)	79.49 % (82.54)
5	Относительные интенсивности + асимметрия и эксцесс (20)	76.00 % (86.67)	90.32 % (89.66)	100 % (100)	88.46 % (92.06)
	Для полного числа лазерных линий (405, 458, 476, 488, 496, 514, 543 нм)				
6	Асимметрия и эксцесс (14)	64.00 % (73.33)	80.65 % (82.76)	100 % (100)	80.77 % (85.71)
7	Отношения интенсивностей (14)	76.00 % (93.33)	77.41 % (86.21)	100 % (100)	83.33 % (92.06)
8	Относительные интенсивности (21)	84.00 % (86.67)	90.32 % (93.10)	100 % (100)	91.03 % (93.65)
9	Отношения интенсивностей + асимметрия и эксцесс (28)	96.00 % (93.33)	87.10 % (96.55)	100 % (100)	93.59 % (96.83)
10	Относительные интенсивности + асимметрия и эксцесс (35)	88.00 % (100)	93.55 % (100)	100 % (100)	93.59 % (100)

В табл. 2 представлены расчетные данные по процентам правильности определения всех трех штаммов для рассматриваемых наблюдений при различных комбинациях параметров. Во второй колонке в скобках приведено количество параметров, используемых для классификации. В остальных колонках в скобках приведены значения процента правильности классификации для выборок, из которых исключены наблюдения, полученные с клеток в измененном физиологическом состоянии (общая выборка 63 наблюдения).

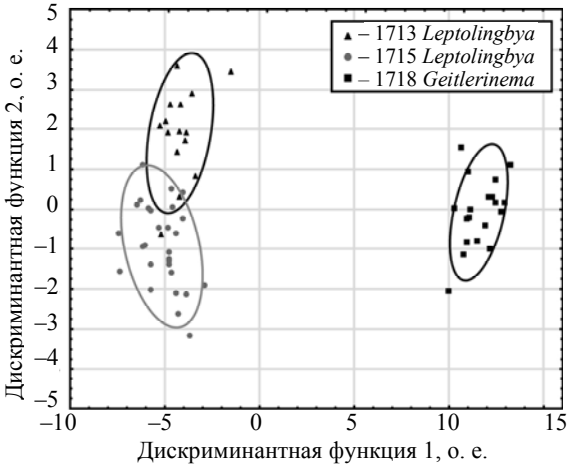
Как следует из приведенных результатов, из двух групп параметров, непосредственно описывающих пигментный состав (2 и 3), лучшее разделение по штаммам дает группа «относительные интенсивности». Это объясняется тем, что, с одной стороны, для расчета этой группы параметров производится усреднение по некоторой области, что уменьшает влияние случайных отклонений в исходных данных. А с другой стороны, при расчете процентного вклада каждого пигмента в общую интенсивность флуоресценции учитываются не только максимальные значения в пиковых точках, но и общая форма кривой. Добавление к группам 2 и 3 параметров «асимметрии и эксцесса» позволяет увеличить точность классификации на 4…6 % по обеим группам параметров (группы 4 и 5), хотя в отдельности группа 1 дает достаточно низкий результат разделения штаммов. Это объясняется тем, что данная группа описывает форму спектральной линии в целом и позволяет учитывать такие характеристики, как расстояние между пиками и ширина пиков, а следовательно, более точно описывает все характерные особенности спектров одного рода. Из приведенных данных следует, что к классу 1718 все наблюдения отнесены верно, так как спектры флуоресценции для данного штамма сильнее всего отличаются от двух других в данной выборке. Ошибочно классифицируются наблюдения для близкородственных штаммов 1713 и 1715, т. е. основная ошибка возникает при внутриродовом разделении. Поскольку эти штаммы филогенетически и морфологически очень близки, данная ошибка считается допустимой.

Использование исключенных данных по другим лазерным линиям, а также исключение из выборки наблюдений, полученных с клеток в измененном физиологическом состоянии, повышают процент правильности классификации на 8…14 %. На рис. 3 представлены результаты дискриминантного анализа при совместном применении групп параметров «асимметрия и эксцесс» и «относительные интенсивности»: рис. 3, *а* – при включенных наблюдениях, полученных с клеток в измененном физиологическом состоянии, и при применении ограниченного количества лазерных линий (классификация по 36 параметрам); *б* – при исключении наблюдений, полученных с клеток в измененном физиологическом состоянии, и при классификации по 36 параметрам; *в* – при исключении наблюдений, полученных с клеток в измененном физиологическом состоянии, и при применении семи лазерных линий (классификация по 35 параметрам). Наблюдается очевидное возрастание точности классификации.

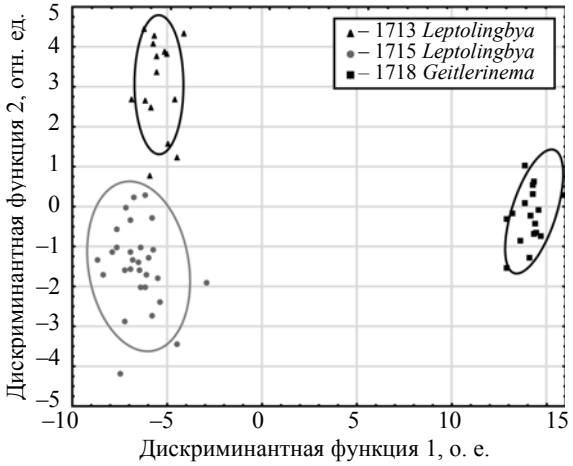
Для проверки правильности классификации и устойчивости анализа при уменьшении выборки было проведено следующее исследование. Из начальной выборки для каждого штамма, содержащей экспериментальные наблюдения, полу-



а



б



в

Рис. 3

Таблица 3

Штамм (дата наблюдения)	Квадраты расстояния Махаланобиса (d_M)		
	1713	1715	1718
1713 *Leptolyngbya* (08.02.17, 24.03.17)	11.5523	20.07029	244.0343
1715 *Leptolyngbya* (24.03.17, 29.03.17)	13.26994	14.75548	280.4088
1718 *Geitlerinema* (08.02.17)	309.4949	345.7305	8.98072

ченные в разные дни, была выделена некоторая часть наблюдений за определенную дату, которую для данного исследования не приписывали ни к одному из определенных классов и считали неизвестной. Далее, после проведения дискриминантного анализа на частичных выборках выделенные выборки были классифицированы при помощи полученных дискриминантных функций. Результаты представлены на рис. 4: жирные линии – частичные выборки; тонкие линии – наблюдения за конкретную дату (с изначально неопределенной классификацией). Расстояния Махаланобиса между частичными и выделенными выборками (табл. 3) показывают, что выборки за определенные даты находятся ближе всего к классу, к которому они должны принадлежать. Таким образом, при выделении части выборки в отдельный класс дискриминантный анализ на основе оптимального набора параметров (строка 5 в табл. 2) правильно классифицирует неизвестные штаммы.

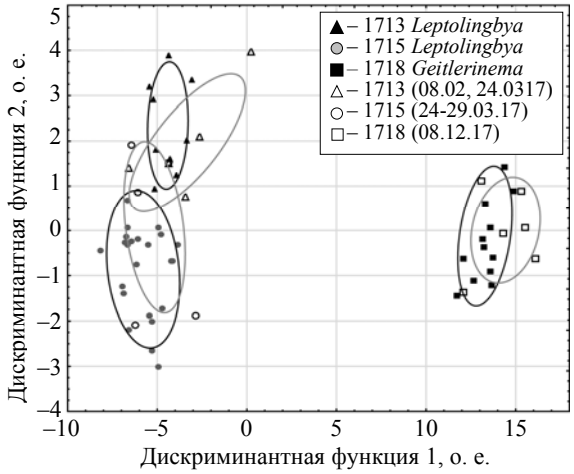


Рис. 4

Следует отметить, что исключенные в данном исследовании малозначимые параметры для других групп наблюдений могут оказаться решающими при дискриминации. Так, например, в данном случае на первом этапе была исключена из рассмотрения первая спектральная область, однако для спектров флуоресценции цианобактерий, содержащих фикоэритрин в качестве основного пигмента светособирающего комплекса, в данной области спектра находится дополнительный пик, определяющий их родовую принадлежность. Таким образом, в каждом отдельном случае, для каждой выборки следует отдельно проводить анализ параметров на их значимость для классификации.

Впервые на основе дискриминантного анализа была проведена дифференциация родов цианобактерий. Разработанная методика получения и обработки спектров собственной флуоресценции отдельных клеток позволила провести дифференциацию даже близкородственных штаммов с достаточно высокой точностью. В результате исследования было показано, что при правильной селекции и обработке исходных данных спектры собственной флуоресценции отдельных живых клеток дают достаточно статистически достоверной информации для проведения классификации до рода/штамма цианобактерий. Очевидно, что представленная методика извлечения параметров из спектров собственной флуоресценции отдельных клеток позволяет автоматизировать процедуру первичной дифференциации родов цианобактерий в натурных пробах и оптимизировать процесс непрерывного экологического мониторинга открытых водоемов. Описанная в данной статье возможность значительного уменьшения количества необходимых параметров классификации за счет исключения наименее значимых из них является ключевым фактором при создании автоматизированного программно-аппаратного комплекса для экологического мониторинга, поскольку уменьшение числа возбуждающих линий вдвое без потери эффективности межродовой классификации может значительно упростить и удешевить прибор. Кроме того, разработанная в среде MATLAB компьютерная программа рассчитана на любое количество исходных данных и уже была успешно протестирована при классификации 21 штамма цианобактерий по 120 параметрам, извлеченным из спектров собственной флуоресценции.

Все экспериментальные данные были получены с использованием оборудования РЦ «Развитие молекулярных и клеточных технологий» Научного парка СПбГУ. Образцы штаммов цианобактерий предоставлены РЦ «Культивирование микроорганизмов» Научного парка СПбГУ.

СПИСОК ЛИТЕРАТУРЫ

1. High-resolution phytoplankton diel variations in the summer stratified central Yellow Sea / X. Liu, B. Huang, Z. Liu, L. Wang, H. Wei, C. Li, O. Huang // J. of oceanography. 2012. Vol. 68, № 6. P. 913–927.

2. CHEMTAX – a program for estimating class abundances from chemical markers: application to HPLC measurements of phytoplankton / M. D. Mackey, D. J. Mackey, H. W. Higgins, S. W. Wright // Marine Ecology Progress Series. 1996. Vol. 144. P. 265–283.

3. In-vivo absorption characteristics in 10 classes of bloom-forming phytoplankton-taxonomic characteristics and responses to photoadaptation by means of discriminant and HPLC analysis / G. Johnsen, O. Samset, L. Granskog, E. Sakshaug // Marine Ecology Progress Series. 1994. Vol. 105, № 1–2. P. 149–157.

4. Using absorbance and fluorescence spectra to discriminate microalgae / D. F. Millie, O. M. Schofield, G. J. Kirkpatrick, G. Johnsen, T. J. Evens // European J. of Phycology. 2002. Vol. 37, № 3. P. 313–322.

5. Quantifying phytoplankton communities using spectral fluorescence: the effects of species composition and physiological state / N. Escoffier, C. Bernard, S. Hamlaoui, A. Groleau, A. Catherine // J. of Plankton Research. 2014. Vol. 37, № 1. P. 233–247.

6. A fluorometric method for the differentiation of algal populations in vivo and in situ / M. Beutler, K. H. Wiltshire, B. Meyer, C. Moldaenke, C. Luring, M. Meyerhofer // Photosynthesis research. 2002. Vol. 72, № 1. P. 39–53.

7. MacIntyre H. L., Lawrenz E., Richardson T. L. Taxonomic discrimination of phytoplankton by spectral fluorescence // Chlorophyll a fluorescence in aquatic sciences: methods and applications. Dordrecht: Springer, 2010. P. 129–169.

8. Identifying phytoplankton in seawater based on discrete excitation-emission fluorescence spectra / F. Zhang, R. Su, J. He, M. Cai, W. Luo, X. Wang // J. of phycology. 2010. Vol. 46, № 2. P. 403–411.

9. Фаткуллина И. Б., Протопопова Н. В., Михалевич И. М. Дискриминантный анализ как метод проведения дифференциальной диагностики артериальной гипертензии при беременности // Вестн. новых медицинских технологий. 2011. Т. 18, № 1. С. 1–2.

10. Fisher R. A. The Use of Multiple Measurements in Taxonomic Problems // Annals of Eugenics. 1936. Vol. 7. P. 179–188.

11. Hartigan J. A., Wong M. A. Algorithm AS 136: A k-means clustering algorithm // J. of the Royal Statistical

Society. Series C (Applied Statistics). 1979. Vol. 28, № 1. P. 100–108.

12. Corpet F. Multiple sequence alignment with hierarchical clustering // Nucleic acids research. 1988. Vol. 16, № 22. P. 10881–10890.

13. Hansen M., Dubayah R., DeFries R. Classification trees: an alternative to traditional land cover classifiers // Intern. J. of remote sensing. 1996. Vol. 17, № 5. P. 1075–1081.

14. Langley P., Sage S. Induction of selective Bayesian classifiers // Proc. of the Tenth Intern. Conf. on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc. Seattle: WA, 1994. P. 399–406.

T. R. Jhangirov, A. S. Perkov, A. A. Liss
Saint Petersburg Electrotechnical University «LETI»

N. Yu. Grigoryeva
Saint Petersburg State University

L. V. Chistyakova
Center for Culture Collection of Microorganisms SPSUof Saint Petersburg State University

APPLICATION OF LINEAR DISCRIMINANT ANALYSIS FOR CLASSIFICATION OF CYANOBACTERIA BY INTRINSIC FLUORESCENCE SPECTRA

A technique for cyanobacterial species classification by in-vivo single-cell fluorescence spectra was elaborated using special methods of mathematical statistics. The analysis of several hundred spectra for 20 different cyanobacterial strains, obtained by means of confocal laser scanning microscopy, was carried out. A specific procedure for input data processing and the order of extraction of such parameters as fluorescence spectra shape and the ratio of individual peaks were elaborated. Statistical methods were used for determination of a limited set of key parameters sufficient for classification. To solve the classification problem a standard multivariate discriminant analysis was used. As an example, the proposed classification algorithm was applied for the differentiation of three cyanobacterial strains, belonging to two genera.

Cyanobacteria, discriminant analysis, statistical methods for biology, fluorescence spectra

УДК 681.3

Н. А. Верзун, М. О. Колбанев, Б. Я. Советов, А. И. Яшин
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В. И. Ульянова (Ленина)

Методы сбора данных с сенсорных узлов беспроводной сенсорной сети

Рассматривается процесс информационного взаимодействия сенсорных узлов интернета вещей с базовой станции. Сформулированы требования к устройствам и алгоритмам. Представлена классификация методов опроса узлов сенсорного поля, основанных на процедурах случайного доступа к радиоканалу. Описаны две модификации случайного метода доступа без контроля несущей: модифицированный синхронный случайный и случайный синхронно-временной доступ к радиоканалу, которые по сравнению с известными методами ведут к уменьшению числа коллизий в системе сбора данных. Для каждого типа доступа приведены алгоритмы и временные диаграммы, представляющие процесс сбора базовой станцией данных с сенсорных узлов. Выявлена задача разработки новых алгоритмов взаимодействия устройств в беспроводных сенсорных сетях в качестве связующего звена между сенсорными узлами (физическими умными вещами) и соответствующим им облачным ресурсом Интернета. Задача уменьшения потребления энергии на всех этапах работы сенсорных узлов признана первоочередной и позволяет продлить срок службы сенсорной сети в целом. Для организации сбора данных с сенсорных узлов беспроводных сенсорных сетей целесообразно применять случайные «традиционные» методы доступа к радиоканалу, которые необходимо модифицировать с целью уменьшения числа коллизий в беспроводных сенсорных сетях.

Интернет вещей, сенсорное поле, сенсорный узел, базовая станция, сбор данных, методы доступа к радиоканалу, синхронный случайный доступ, модифицированный синхронный случайный доступ, случайный синхронно-временной доступ

Технологической основой концепции интернета вещей являются беспроводные сенсорные сети (БСС). БСС представляет собой самоорганизующуюся структуру [1], [2], которая обеспечивает информационное взаимодействие огромного числа сенсорных узлов (СУ), входящих в ее со-