УДК 579.8.06

КОНЦЕПЦИЯ УНИВЕРСАЛЬНОЙ ТАКСОНОМИЧЕСКОЙ СИСТЕМЫ БАКТЕРИЙ: ЭВОЛЮЦИОННОЕ ПРОСТРАНСТВО ГЕНА 16S-рРНК v. 1.0*

А.С. ДОЛЬНИК[1], Г.С. ТАМАЗЯН[1], Е.В. ПЕРШИНА[2], К.В. ВЯТКИНА[3], Ю.Б. ПОРОЗОВ[4], А.Г. ПИНАЕВ[2], Е.Е. АНДРОНОВ[2]

Проблема системности в таксономии, в основе своей связанная с вопросами эволюции, остается одной из сложнейших в современной биологии, и в частности в микробиологии. Эта проблема всегда привлекала внимание ученых, в том числе Н.И. Вавилова, закон гомологических рядов которого, несомненно, следует отнести к числу наиболее ярких попыток внесения упорядоченности в анализ биоразнообразия. В молекулярной экологии микроорганизмов востребованность универсальной таксономической системы особенно очевидна. Анализ таксономической структуры почвенных микробиомов с использованием секвенаторов нового поколения сталкивается с многочисленными трудностями, одна из которых — невозможность точной идентификации значительной части выявляемых в окружающей среде вариантов гена 16S-рРНК из-за отсутствия близкородственных последовательностей в базах данных. Для решения этой проблемы мы предлагаем концепцию «эволюционного пространства» гена 16S-рРНК, или своеобразной системы, в которой есть место для любой последовательности указанного гена вне зависимости от того, присутствует ли она в базах данных/биосфере и даже реализована ли она в ходе эволюции. В такой системе любой вариант гена 16S-рРНК получает фиксированные координаты. Эволюционное пространство открывает возможность для создания универсальной «таксономической карты» и привлечения ряда мощных алгоритмов для анализа микробного сообщества как единого целого. В настоящей публикации описана первая версия эволюционного пространства гена 16S-рРНК бактерий минимально возможной размерности (13D).

Ключевые слова: эволюционное пространство, метатаксономия, 16S-рРНК, микробное сообщество.

Keywords: evolutionary space, metataxonomy, 16S rRNA, microbial community.

Таксономическая структура биоты, основные закономерности, определяющие эколого-географическое распространение видов и их эволюцию, остаются фундаментальными проблемами биологии со времен выдающихся естествоиспытателей и основателей генетики. По мере накопления научных знаний такие исследования углублялись, достигнув в последние десятилетия уровня оценки молекулярных особенностей генома объектов.

Знаменитый принцип одного из основоположников экологии микроорганизмов М. Бейеринка (M.W. Beyerinck), согласно которому все есть везде, но среда отбирает («Everything is everywhere but the environment selects»), сформулированный 100 лет назад, до сих пор остается весьма плодотворной научной идеей, обладающей мощным эвристическим потенциалом (1). Применительно к сельскохозяйственной микробиологии он указывает на возможность решения обратной задачи — анализа агроэкологического состояния почвы по данным изучения ее микробиома. Настоящая публикация имеет непосредственное отношение именно к этой проблеме. Современный подход к анализу таксономической структуры почвенных микробиомов предполагает выделение почвенной ДНК (РНК), конструирование библиотек таксономически значимых генов (например, 16S-рРНК) и их секвенирование с последующей таксономической идентификацией (2). В результате таких исследований формируются списки выявленных таксонов. С введением в практику молекулярной экологии секвенаторов нового поколения число нуклеотидных последовательностей, характеризующих

* Работа поддержана Министерством науки и образования РФ (ГК № 16.552.11.7047) и Программой поддержки фундаментальных исследований по приоритетным направлениям Санкт-Петербургского государственного университета.

отдельный образец, достигло десятков тысяч (3, 4). Работать со столь обширными списками достаточно сложно, несмотря на обилие программного обеспечения (5-7), созданного для этих целей. К важным и не разрешенным в настоящее время проблемам, связанным с анализом подобных данных, относятся сложность учета таксономически не атрибутируемых последовательностей, принадлежащих, как правило, к еще не описанным таксонам, которые нередко составляют значительную долю микробных сообществ в окружающее среде; затрудненность одновременного анализа нуклеотидных последовательностей, представляющих различные участки одного и того же гена; отсутствие интегральных статистических подходов, позволяющих описывать сложные сообщества как единое целое.

Для решения этих проблем нами была сформулирована задача построения «эволюционного пространства» для описания глобального нуклеотидного разнообразия и эволюционных процессов в пределах одного единственного гена, в частности гена 16S-рРНК. Определение такого пространства весьма лапидарно: эволюционное пространство гена 16S-рРНК — это метрическое пространство, в котором нуклеотидные последовательности гена, принадлежащего различным микроорганизмам, представлены точками, а расстояния между каждой парой точек отражают эволюционные расстояния между соответствующими нуклеотидными последовательностями. Подобная задача традиционно относится к области многомерного шкалирования и практически сводится к размещению матрицы попарных генетических расстояний в метрическом пространстве таким образом, чтобы геометрические расстояния между точками соответствовали эволюционным дистанциям. При этом подразумевается, что для любой нуклеотидной последовательности гена 16S-рРНК могут быть вычислены геометрические координаты, характеризующие положение точки в пространстве. Между тем за столь простым определением скрывается чрезвычайно широкий круг проблем из области молекулярной эволюции, таксономии, математики, геометрии, к тому же вычислительная сложность подобной задачи очень высока. Но в случае успеха такое пространство могло бы стать принципиально новой операционной средой для анализа данных молекулярного таксономического анализа сложных микробных сообществ с введением ряда новых интегральных характеристик, таких как плотность, объем, геометрия, центральная точка (сообщества в целом или отдельных таксонов) и т.д. Кроме того, представители любого, даже неизвестного таксона, получают фиксированное положение, что чрезвычайно облегчает анализ неатрибутируемых компонентов микробного сообщества и открывает возможности для создания универсальной «таксономической карты», в которой для любой последовательности найдется закрепленное за ней место.

Интересно, что обсуждаемая проблема, по всей видимости, тесно связана с одной из фундаментальных в таксономии — построением так называемой естественной классификации организмов, в которой каждый объект занимает положение, соответствующее его родству с другими организмами. Впервые она была ясно сформулирована еще К. Линнеем в труде «Философия ботаники» в 1751 году, однако не решена до сих пор (8). Большое внимание этой проблеме уделял Ч. Дарвин (9). Из отечественных ученых в первую очередь следует упомянуть Н.И. Вавилова (10) — его вклад в концепцию биологического вида и открытый им закон гомологических рядов, устанавливающий параллелизм в наследственной изменчивости организмов и, безусловно, представляющий собой одну из пионерских работ по созданию единой системы, которая не только описывает наличествующее биоразнообразие, но и указывает на отсутствующие таксоны.

Обсуждение истории вопроса выходит далеко за рамки настоящего исследования, и построение такой системы не составляет его цели. Наша задача сводится к описанию разнообразия нуклеотидных последовательностей одного лишь бактериального гена 16S-рРНК с использованием представления об эволюционном пространстве.

Методика. Для анализа использовали релиз SSURef_104_SILVA_NR_99, доступный на сервере SILVA (http://www.arb-silva.de/no_cache/download/archive/release_102/Exports/), содержащий выровненные нуклеотидные последовательности гена 16S-рРНК высокого качества длиной не менее 1000 н., причем в этом релизе удалены все последовательности со сходством более 99 %, что, со всей очевидностью, не может повлиять на геометрические соотношения. После исключения архей в релизе осталась 210 651 нуклеотидная последовательность, соответствующая бактериальным генам 16S-рРНК. Инструментами и программными средствами служили <РСУБД MSSQL Server 2008 R>, MatLab R2009b, Revolution R 4.3, индивидуальные программы на c, c#, c++.

Построение матрицы расстояний и поиск симплексов. Для вычисления попарных расстояний между последовательностями использовали p-distance (pairwise deletion), представляющую собой долю различающихся нуклеотидных позиций, вычисляемую при попарном удалении позиций, содержащих пропуски и вырожденные нуклеотиды. На основании вычислений была сформирована матрица попарных расстояний для всей базы данных с идентификаторами последовательностей. Кроме того, вычислили распределение фил по численности в базе данных и распределение попарных расстояний в матрице.

Для поиска симплексов был выбран диапазон расстояний [0,251-0,269], обеспечивающий невозможность ошибочного включения в симплекс радиуса: в бесконечномерном симплексе отношение ребра к радиусу соответствует $\sqrt{2}$ (доказательство не приводится), а в случае 14-мерного пространства оно составляет ~ 1,463 (0,269/1,463 ~ 0,183 << 0,251). Ввиду того, что в разумное время не представляется возможным осуществить полную проверку всех имеющихся вариантов даже в усеченной базе, для поиска симплекса был предложен так называемый жадный алгоритм, основанный на выявлении последовательностей-кандидатов, характеризующихся максимальным числом попарных дистанций, лежащих в заданном диапазоне, с последующим пошаговым расширением списков. С целью расширения области поиска в алгоритм была добавлена стохастическая функция (случайный выбор из списков кандидатов), использование которой оказалось очень эффективным. Результатом вычислений было выявление серии симплексов различного размера, из которых для дальнейшего анализа выбирали максимальные.

Картирование последовательностей. Для картирования последовательностей в эволюционном пространстве был выбран симплекс 6 (см. раздел «Результаты»). Позиционирование точек осуществляли следующим образом: точки симплекса ({s1, s2, ..., s14}) размещали на координатных осях в соответствии с номерами позиций (номер точки соответствует номеру оси). С этой целью было принято одинаковое расстояние между последовательностями — вершинами симплекса (хотя в точности оно таковым не является), нормированное к единице. Таким образом, все расстояния в системе тоже масштабировались пропорционально масштабированию симплекса, а именно делились на среднее арифметическое расстояний между вершинами симплекса, то есть на 0,261.

Нахождение координаты точки для каждого из вариантов гена

16S-рРНК в пространстве осуществлялось следующим образом. Изначально имеются расстояния от последовательности X до опорных последовательностей — вершин симплекса {r1, r2, ..., r14}, нормированные на приведенный выше коэффициент. Ищем такие 14 точек {x1, x2, .., x14} в нашем пространстве, для которых выполняются два условия: первое — точки находятся на соответствующем расстоянии от вершин симплекса, то есть dist(x1, s1) = r1, dist(x2, s2) = r2, ..., dist(x14, s14) = r14; второе — суммарное расстояние (штрафная функция) между этими точками минимальное: dist(x1, x2) + dist(x1, x3) + ... + dist(x13, x14) → min.

Эта задача относится к задачам квадратичной (нелинейной) оптимизации с граничными условиями. Для решения задач подобного рода существует быстро сходящийся метод градиентного спуска с множителями Лапласа, однако в нашем случае, поскольку граничные условия квадратичны, оптимизация проводилась с использованием алгоритма interior point в реализации функции fmincon из Optimization Toolbox MatLab.

После нахождения множества точек {x1, x2, .., x14} приведенным выше методом за координату точки X принимается центр масс (покоординатное среднее арифметическое) найденных точек. В случае если найденный центр масс не попадает на выбранный симплекс, применяется ортогональная проекция. Следует отметить, что по результатам эксперимента значение штрафной функции не сильно отличалось от нуля, что указывает на близость полученных точек {x1, x2, .., x14} друг к другу. Геометрические расстояния функции dist вычислялись в соответствии с евклидовой метрикой.

Для оценки точности картирования рассчитывали корреляции между матрицами попарных расстояний — истинной и вычисленной по геометрическим координатам с использованием метода Мантеля (11).

Визуализация распределений. Визуализацию распределений точек в 14-мерном пространстве (на самом деле мы имеем дело с 13-мерным построением, дополнительная размерность была введена лишь для удобства вычислений) проводили при помощи построения срезов двумерными плоскостями с небольшой толщиной, заданной таким образом, чтобы она попадала в диапазон, который и так нельзя различить ввиду ошибок округления в вычислениях и дискретности расстояния p-distance. Всего выполнили около 1000 сечений сериями параллельных плоскостей, выбранных по одному из направлений и обозначенных двумя осями, через которые проходит базовая плоскость (например, 2-12 — плоскость, на которой лежат оси 2 и 12).

Результаты. Идея отображения нуклеотидных последовательностей гена 16S-рРНК в виде точек в пространстве не нова. Однако большинство исследователей в этой области оперируют со статистическими подходами, ориентированными на построение различного рода проекций, например методами главных компонент, переводящими матрицу попарных расстояний в пространственные отображения (12, 13). Наиболее близка к поставленной задаче попытка построения многомерного векторного пространства для отображения эволюционного процесса у позвоночных по данным аминокислотной последовательности α-гемоглобина (14). Основной идеей предлагаемого нами практического подхода к реализации этой задачи также было представление о многомерности эволюционного пространства и принципиальной невозможности таких построений в пространствах малой размерности. В настоящем построении опорным элементом были симплексы — геометрические фигуры, у которых расстояния между двумя любыми вершинами одинаковы. Примерами простейших симплексов служат равно-

сторонний треугольник (2D) и правильный тетраэдр (3D). Симплексы с 5 и более вершинами также существуют, но могут быть построены только в многомерных пространствах. Поиск симплексов мы осуществляли в матрице попарных эволюционных дистанций одного из релизов международной базы данных по разнообразию гена 16S-рРНК. Выявленные симплексы определили минимальную размерность целевого пространства и были использованы для дальнейшего картирования в качестве реперных точек.

Структура базы данных и особенности распределения попарных расстояний. На рисунке 1 приведены распределения фил по численности в рабочей базе (отображены только филы с этим показателем более 0,9 % от общего числа записей) и попарных расстояний в соответствующей матрице.

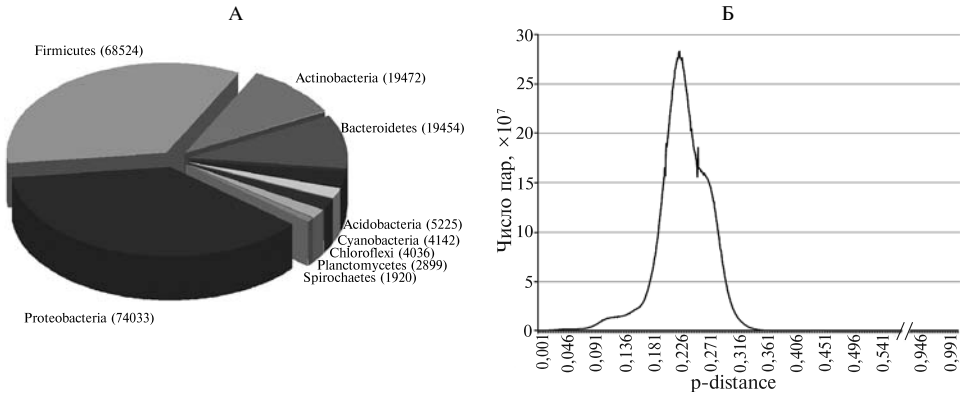

Рис. 1. Распределение нуклеотидных последовательностей гена 16S-рРНК в рабочей базе данных по филам микроорганизмов (А) **и распределение попарных расстояний в построенной матрице** (Б).

Следует отметить весьма неравномерное распределение фил в базе данных, что несколько затрудняет анализ. Основная масса записей относится к филам *Proteobacteria*, *Firmicutes*, *Actinobacteria*, *Bacteroidetes*, *Acidobacteria*, *Cyanobacteria*, *Chloroflexi*, *Planctomycetes*, *Spirochaetes*, *Verrucomicrobia*, *Gemmatimonadetes*, *Nitrospirae*, *Lentisphaerae*, *Synergistetes*, *Chlorobi*. Для *Chlorobi*, наименее представленной из приведенных фил, имеется 505 записей. Такое распределение обусловлено, во-первых, неравномерностью распространении фил в биосфере, во-вторых, структурой и целями тех научных изысканий, в результате которых происходит наполнение баз данных. Понятно, что для предпринимаемых в настоящем исследовании вычислений оптимальным было бы равномерное распределение, однако на текущий момент именно использованная база данных — одна из наиболее полных, поэтому с подобным отклонением приходится мириться.

Выявление симплексов. В результате проведенного анализа в базе данных было выявлено 25 приблизительно правильных симплексов с 14 вершинами каждый, соответствующие 13-мерному пространству, при расстоянии между вершинами в диапазоне [0,251-0,269]. Очевидно, что вряд ли возможно найти в базе данных абсолютно правильные симплексы, поэтому при вычислениях мы исходили из диапазона, ширину которого выбрали так, чтобы исключить попадание в симплекс радиуса (см. раздел «Методика»). Ниже приведен список выявленных симплексов и графическое представление их распределения среди основных бактериальных фил (рис. 2):

1-й симплекс EU773611; EU491566; AJ542543; AY485285; AB355037; X86688; EU703430; Y10649; EF096697; EF516823; EU804917; AY571792; AM420109; AJ306801

2-й симплекс EU469976; EU503653; GQ502583; AF189244; AY212563; AY907749; GQ397076; FJ628180; DQ814080; EU669608; DQ811945; AB191897; GQ346956; EU245865

3-й симплекс FJ231137; EU135237; DQ795973; EU776122; AY863081; EU881151; EF020301; EU802835; AB488334; AB300126; EU038002; EU246179; FJ545465; CU924649

4-й симплекс	AF419696; EU506479; EU507872; DQ811928; D11348; FJ456773; X71862; FN563192; CP001110; FJ648694; AF068427; EU335420; AY743263; FN556062
5-й симплекс	EU381735; EF688230; EU370505; EF454921; EU799550; EF575061; FJ821610; FN401325; GU061319; FJ873298; AY280413; EU135375; FJ592895; GQ350871
6-й симплекс	FJ438004; GQ246374; FJ881166; EU005687; X73976; EU463251; DQ337095; AY225654; AY605151; FJ478836; FJ628268; FJ901103; CP001080; CU925754
7-й симплекс	DQ803694; EU767531; X12742; X81063; EU869405; EU134585; EU360497; AY571796; AY197394; AB177131; EF203193; FJ976270; EU134048; FJ891053
8-й симплекс	EU465688; EU511290; FJ366892; AB188635; AY663886; GU127275; EU775151; FJ717259; EU804722; FJ456653; EU491403; AJ431238; CU922689; FJ516821
9-й симплекс	EF575007; FJ748813; EU366375; AM712329; DQ248296; FJ983028; GQ263308; FJ802296; AY605160; EF076074; DQ906017; AB294345; CU923425; DQ330595
10-й симплекс	EU074225; CT573820; CU925797; DQ800076; EU037954; FJ976253; AB464934; DQ308543; FJ192842; EF019248; EU250258; AB243263; EU133963; X84212
11-й симплекс	AB277853; EU478629; EF522262; EU772741; EU635952; AY907749; CP001099; EU134803; EU159562; AB245338; GQ397047; AB192244; CU923893; DQ499300
12-й симплекс	EU507587; U32593; M24483; AM712329; FJ826329; AJ867904; EF019021; AF543503; DQ676428; EU266879; CU922282; EU043840; GQ340131; DQ906038
13-й симплекс	EU505590; FJ858737; FJ628297; AB240485; EU134568; U91515; EU132320; AB031999; CU921210; EU134128; GQ264185; EU289449; CU920242; FJ625343
14-й симплекс	FJ748815; EU503864; FJ159133; EU010170; EF453815; EU135522; EF018434; EF515949; X86774; AB198654; CU918198; AB462555; CR933027; FJ264554
15-й симплекс	AY114316; EU463474; GQ275102; FJ382145; AY672075; DQ906842; DQ005880; GQ264171; M79383; EU134919; AM934777; EU134038; CU921544; EU850520
16-й симплекс	EF520637; FJ873260; DQ809643; EU617874; AB286524; EU470375; FN430655; FJ478875; DQ811949; EF203193; EU245088; FJ478622; CU925754; EF192905
17-й симплекс	AB192054; EU509270; GQ441271; AY188316; DQ906842; AB286350; GU118530; AY945884; AB088905; CU922949; FJ517055; CU918272; CU927871; AY114333
18-й симплекс	EU778001; EU763449; AY726960; AB355083; FJ592715; FJ516977; EF190824; AY947962; CP000814; EU132011; FJ712505; EU592424; EU134203; FM873402
19-й симплекс	AF317763; EU459226; EU639371; FJ002234; EF592610; EF205470; EU133431; FJ167503; AY913233; AB198604; CU924983; EU662508; FJ712493; AB282966
20-й симплекс	EF019165; EU982406; FJ425646; FN563173; DQ383304; AY349381; EU134307; AF093251; CU918643; CU924912; EU133993; EU247889; EU245649; EU885068
21-й симплекс	AB192219; D11348; EU802784; CU923009; AJ291826; EU773650; GU061962; EU334768; FJ592772; EU385703; AY571473; FJ825446; AB465709; FJ004754
22-й симплекс	AB302409; EU669636; EU434533; FJ493498; FJ790619; FJ746187; EF379616; CU921631; EU915265; AF393378; DQ676384; AB234287; AB525461; AB089051
23-й симплекс	AY862537; EF097759; EU775762; GQ487946; AF385521; AJ306807; EU050858; EU802639; AY913288; EU236294; X89045; CU925964; AF521187; AB294345
24-й симплекс	DQ015655; EU507714; FJ425597; CP001739; FJ802178; FJ985790; FJ628291; EU409852; GQ402806; EF688228; EU721768; CU926616; DQ988318; GQ249498
25-й симплекс	FN554390; EU074225; AY266450; AB034054; FJ002173; EF522341; AJ299413; FJ628241; GQ355003; AF402980; CU927201; EU181504; AB237731; FJ879997

Распределение симплексов по ветвям филогенетического древа продемонстрировало, что эволюционные соотношения между филами гораздо сложнее вытекающих из обычных таксономических представлений. Так, симплекс 6 объединял 14 записей из фил, равномерно распределенные по всему древу, включая крайние группы — *Verrucomicrobia* и *Aquificae*. Выявленные соотношения в распределении симплексов свидетельствовали о том, что размещение имеющегося множества точек с сохранением попарных расстояний невозможно ни в 2D-, ни в 3D-пространствах. Из полученных результатов с очевидностью следует, что задача оценки соотношений между филами неразрешима в пространствах с размерностью меньше 13, и это один из важных итогов настоящего исследования.

Представленность фил в симплексах была приблизительно равномерной, за исключением объектов из фил *Spirochaetes* и *Chloroflexi*, которые встречались примерно в 7 и 5 раз чаще ожидаемого. В остальном представленность фил в симплексах и в базе данных оказалась в общем соизмеримой. Такие широко представленные в базе данных филы, как *Firmicutes* и *Proteobacteria*, характеризующиеся к тому же высокой степенью разнообразия, в пределах симплекса нередко присутствовали более одного раза. То есть генетические дистанции в пределах одной филы могут быть не меньше, чем между филами.

Для поиска симплексов был выбран ограниченный интервал гене-

тических дистанций, обеспечивающий внешнюю локализацию симплекса по отношению к совокупному множеству. По этой причине остается открытым вопрос о максимальном размере симплексов с меньшим расстоянием между вершинами, например внутри филы. Не исключено, что такие симплексы могут иметь большую размерность и эффективное картирование с использованием внешнего симплекса в качестве репера невозможно.

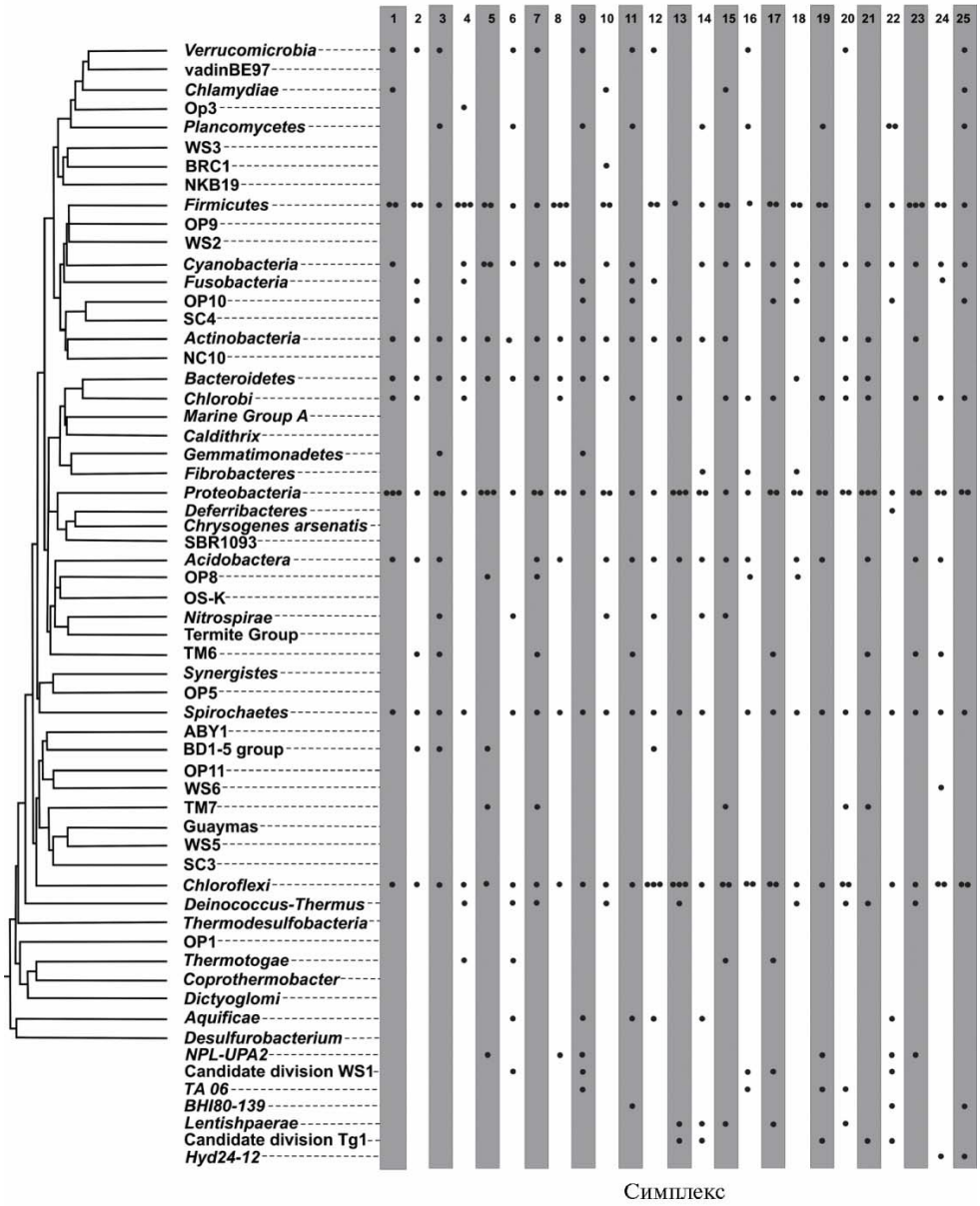

Рис. 2. Распределение 25 выявленных 14-вершинных симплексов среди основных бактериальных фил (на основании нуклеотидных последовательностей гена 16S-рРНК). Число точек в ячейке соответствует числу записей в филе; семь не вошедших в построение дополнительных фил приведены ниже древа. Схематическое представление по M.S. Rappe с соавт. (15) в произвольном порядке.

Картирование нуклеотидных последовательностей в эволюционном пространстве и построение сечений. За реперную основу для картирования последовательностей был выбран симплекс 6, так как в нем каждая из вершин относится только к одной филе и, кроме то-

го, в него включена одна из наиболее удаленных фил — *Aqufiicae.* В соответствии с описанной процедурой для всех последовательностей базы данных в 14-мерном пространстве получили геометрические координаты. На основании данных вычислений была сконструирована восстановленная матрица попарных расстояний. Коэффициент ее корреляции с истинной матрицей (*r*) был невелик и составил 0,299, хотя и имел довольно высокую значимость (односторонний критерий — статистический тест, применяемый для проверки альтернативной статистической гипотезы, выдал значение p = 0,000999). Таким образом, на этом этапе исследований нам не удалось достичь высоких коэффициентов корреляции, что, скорее всего, обусловлено недостаточной размерностью используемого пространства. На следующем этапе исследований нами была предпринята попытка визуализировать распределение точек в пространстве, так как топологические соотношения могут сохраняться и при отсутствии высоких корреляций. Очевидно, что не существует способов прямой визуализации полученного множества, поэтому мы прибегли к способу, аналогичному компьютерной томографии, — созданию плоских сечений (см. раздел «Методика»). На рисунке 3 (А, Б) представлены основные паттерны, выявленные при изучении сечений, показавшем, что в результате было построено эволюционное пространство, характеризующееся низкой, хотя и значимой корреляцией с истинной матрицей попарных расстояний, и демонстрирующее явные тенденции к разделению фил.

В представленной серии срезов, параллельных базовой плоскости 2-12 (см. рис. 3, А), видно последовательное прохождение плоскостью сечения всего множества точек. На полученных сечениях представителям разных фил соответствуют неодинаковые цвета. Очевидно, что наблюдается явная тенденция к разделению бактериальных фил. Более очевидны такие соотношения на серии центральных (то есть проходящих через геометрический центр) срезов (см. рис. 3, Б). Здесь также очевидна четкая тенденция к разделению у представителей разных фил, хотя заметны и зоны смешения, что объясняется, по всей видимости, недостаточной разрешающей способностью пространств небольшой размерности для полной дискриминации таксонов. Тем не менее, некоторые топологические соотношения, выявляемые на срезах, весьма неожиданны. Прежде всего, выполненное построение носит ярко выраженный эволюционный характер. В самом деле, если следовать общепринятой гипотезе о происхождении бактерий от общего предка, то становится очевидным, что эволюционный процесс, представленный в эволюционном пространстве, носит характер расширения, подобного Большому взрыву. Такое расширение необратимо (из статистических соображений) и, по всей видимости, радиально. Следовательно, в этом пространстве существует эволюционный центр (место локализации общего предка), который, скорее всего, должен быть пуст из-за вымывания предковых вариантов гена в ходе глобальной эволюции. Интересно, что существует возможность идентификации подобного центра. Так, по крайней мере две филы — *Proteobacteria* и *Cyanobacteria* имеют явно выраженную вытянутость (см. рис. 3, Б), что не только указывает на высокие скорости эволюции или древность этих фил, но и дает возможность идентифицировать эволюционный центр. В самом деле, если предположение о радиальности расширения верно, то центральные оси, проведенные в данных филах, должны пересекаться (или максимально сближаться) именно в эволюционном центре, и это, безусловно, одно из перспективных направлений исследований. Особый интерес вызывает тот факт, что именно к указанным филам в соответствии с современной так-

сономией относят органеллы эукариотических клеток — хлоропласты (*Cyanobacteria*) и митохондрии (*Proteobacteria*). Интересно также, что по данным предварительного анализа хлоропласты локализованы в дистальной (по отношению к основному массиву бактерий) части филы *Cyanobacteria* (данные не приводятся). Наконец, следует отметить полость, выявленную в пределах филы *Proteobacteria* (см. рис. 3, Б, срез 7-12). Не исключено, что это одна из «эволюционных полостей», наличием которых должны характеризоваться старые монофилетические таксоны, хотя сложность многомерных топологических соотношений не дает возможности утверждать это однозначно.

Вопрос о дальнейшем совершенствовании алгоритма встречается с вполне ожидаемым препятствием, связанным с необходимостью расширения базового симплекса. Мы полагаем, что даже в полной базе данных вряд ли найдется симплекс, заметно больший, чем найдено в настоящем исследовании, так как редукция использованной базы данных была основана на удалении из нее последовательностей со сходством более 99 %. Понятно, что возвращение этих последовательностей в базу не приведет к расширению симплекса, расстояние между вершинами которого соответствует примерно 75 % сходства. Мы предлагаем довольно необычное решение — искусственно сконструировать нуклеотидные последовательности для расширения симплекса (либо de novo, либо посредством коррекции уже существующих записей в базе данных). Помимо технических проблем, имеется ряд вопросов более фундаментального характера, связанных с анализом принципиальной возможности таких построений, их обоснованием и, при необходимости, поиском альтернативных решений.

Итак, еще раз отметим, что цель предпринятого исследования практическая — предложить принципиально новый интегральный подход для анализа многокомпонентных сообществ микроорганизмов в окружающей среде, и прежде всего наиболее сложных почвенных сообществ. В работоспособной версии эволюционного пространства возможно создание универсальной таксономической карты микроорганизмов, в которой фиксированные позиции будут присвоены всем микроорганизмам — таксономически атрибутированным и не атрибутированным, известным и неизвестным. Разработка работоспособной версии эволюционного пространства позволит сформировать предпосылки для введения в анализ данных таких мощных алгоритмов, как, например, распознавание образов, что даст новый импульс в понимании законов формирования микробных сообществ, их эволюции и тонких связей с окружающей средой. Наконец, сама проблема эволюции может раскрыться с весьма неожиданной стороны. Пока же основной задачей остается совершенствование алгоритмов картирования, и именно в этом направлении предполагается развивать начатые исследования.

Выражаем искреннюю признательность В.В. Моттль и В.В. Сулимовой за консультации по ряду вопросов, связанных с представлениями матриц попарных расстояний в метрических пространствах.

ЛИТЕРАТУРА

1. O'Malley M.A. The nineteenth century roots of «everything is everywhere». Nat. Rev. Microbiol., 2007, 5: 647-651.
2. Pace N.R. A molecular view of microbial diversity and the biosphere. Science, 1997, 276: 734-740.
3. Sogin M.L., Morrison H.G., Huber J.A., Mark Welch D., Huse S.M., Phillip R., Neal P.R., Arrieta J.M., Herndl G.J. Microbial diversity in the deep sea and the underexplored «rare biosphere». PNAS USA, 2006, 103: 12115-12120.
4. Tringe S.G., Hugenholtz P. A renaissance for the pioneering 16S rRNA gene. Curr. Opin. Microbiol., 2008, 11: 442-446.

5. Лукашов В.В. Молекулярная эволюция и филогенетический анализ. М., 2009.
6. Kunin V., Copeland A., Lapidus A., Mavromatis K., Hugenholtz P. A Bioinformatician's Guide to Metagenomics. Microbiol. Mol. Biol. Rev., 2008, 72(4): 557-578.
7. Reisenfield S.C., Schloss P.D., Handelsman J. Metagenomics: Genomic analysis of microbial communities. Annu. Rev. Genet., 2004, 13: 525-552.
8. Линней К. Философия ботаники. М., 1989.
9. Дарвин Ч. Происхождение видов путем естественного отбора или сохранение благоприятных рас в борьбе за жизнь. СПб, 1991.
10. Вавилов Н.И. Закон гомологических рядов в наследственной изменчивости. В кн.: Теоретические основы селекции растений. Т. 1. Общая селекция растений /Под ред. Н.И. Вавилова. М.-Л., 1935: 75-128.
11. Mantel N., Valand R.S. A technique of nonparametric multivariate analysis. Biometrics, 1970, 26: 547-558.
12. Garrity G.M., Lilburn T.G. Mapping taxonomic space: an overview of the road map to the second edition of Bergey's Manual of Systematic Bacteriology. WFCC News, 2002, 35: 5-15.
13. Lee S.H., Hwang K.S., Lee H.R et al. Embedding operational taxonomic units in three-dimensional space for evolutionary distance relationship in phylogenetic analysis. Proc. 5th WSEAS Int. Conf. on circuits, systems, electronics, control and signal processing. USA, 2006: 192-196.
14. Kitazoe Y., Kishino H., Okabayashi T., Watabe T., Nakajima N., Okuhara Y., Kurihara Y. Multidimentional vector space representation for convergent evolution and molecular phylogeny. Mol. Biol. Evol., 2004, 22(3): 704-715.
15. Rappe M.S., Giovannoni S.J.. The uncultured microbial majority. Annu. Rev. Microbiol., 2003, 57: 369-394.

[1]Санкт-Петербургский государственный университет,
199034 г. Санкт-Петербург, Университетская наб., 7-9,
e-mail: alexander.dolnik@gmail.com;
[2]ГНУ Всероссийский НИИ сельскохозяйственной микробиологии Россельхозакадемии,
196608 г. Санкт-Петербург—Пушкин, ш. Подбельского, 3,
e-mail: eeandr@gmail.com;
[3]Санкт-Петербургский академический университет, научно-образовательный центр нанотехнологий РАН,
195220 г. Санкт-Петербург, ул. Хлопина, 8, корп. 3,
e-mail: kira@math.spbu.ru;
[4]Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики,
197101 г. Санкт-Петербург, Кронверкский просп., 49,
e-mail: porozov@ifc.cnr.it

Поступила в редакцию 25 мая 2012 года

THE EVOLUTIONARY SPACE OF BACTERIAL 16S rRNA GENE v. 1.0.

A.S. Dolnik[1], G.S. Tamazyan[1], E.V. Pershina[2], K.V. Vyatkina[3], Yu.B. Porozov[4], A.G. Pinaev[2], E.E. Andronov[2]

Summary

A systematicity in taxonomy, basically related to evolution, remains one of the greatest problem of in modern biology, and in particular microbiological topology. This problem has always attracted the attention of scientists, including N.I. Vavilov. He proposed a law of homologous series which, of course, must be regarded as the most striking in the current attempts to make analysis of biodiversity. In the molecular ecology of microorganisms, the demand for universal taxonomic system is particularly evident. Introduction of the new generation sequencing techniques into molecular ecology studies requires introduction of the radically new statistical approaches. This problem can be solved by the creation of the «metataxonomy», an integral approach for the analysis of the microbial communities, allowing to study microbial communities as a whole. It is related to a number of questions in evolutionary biology, taxonomy, mathematics, geometry and demands large computing. One of the most important problems is thr detection in 16S rRNA libraries of large amount of taxonomically «not attributed» sequences. To resolve this problem we propose the «evolutionary space» of 16S rRN gene, where fixed coordinates exist for every possible variant of 16S rRNA gene regardless of whether this variant is present in biosphere/database or even implemented in the course of evolution. In the current article we present the results of the analysis of a 16S rRNA gene database, where for the first time we constructed «evolutionary space», the assumed operational environment for «metataxonomy». The evolutionary space makes it possible to use a number of powerful statistical approaches aimed to analyse complex microbial community as a whole. Here we present the first version of evolutionary space with minimal possible dimension (13D).