Ведущим фактором кластеризации триплетных профилей генов 16S РНК бактерий является таксономия их носителей

Садовский М. Г.[1,2,3], Тетерлева А. А.[4], Моргун А. В.[3], Ларионова И. А.[3], Абрамов В. Г.[1]

[1]ФСНКЦ ФМБА России, Красноярск

[2]ИВМ СО РАН, Красноярск

[3]Красноярский государственный медицинский университет

[4]Сибирский федеральный университет, ИФБиБТ

4 октября 2022 г.

Цель работы:

Выявление связи таксономии и триплетного состава генов 16 S РНК бактерий.

Задачи

1. Сформировать базу генетических данных и провести классификацию/кластеризацию.
2. Проверить гипотезу о связи выявляемых кластеров с таксономией генов, попавших в кластер.

Ключевой вопрос:

Никто не открывает Америку — 16S РНК давно используются в филогении.

Вопрос в другом: верно ли, что классификация **без учителя** способна увидеть связь триплетного состава с таксономией?

Актуальность исследования: для чего?

- если такую связь (на «популяционном» уровне) удастся выявить, то тогда наверняка можно построить систему (полу)автоматической диагностики состояния микробиоты;
- указанная диагностика важна и весьма востребована в медицине — в частности, в диагностике ряда невралгических заболеваний (РС, БП, БА, . . .);
- фундаментальная задача: общая теория связи структуры, функции и таксономии генетических систем.

Используемые методы исследования

В работе использовались следующие методы

- Методы линейного статистического анализа:
  - описательная статистика,
  - метод главных компонент;
  - метод динамических ядер (aka K-means).
- Метод упругих карт (нелинейная статистика).

Возможно применение и других методов нелинейной статистики (например, теоретико-графовых); они в данной работе не использовались.

Суть метода упругих карт

Метод упругих карт является методом нелинейного анализа многомерных данных, эффективно позволяющим уменьшить размерность данных (до 2-х) и выявить неоднородности в распределении данных (кластеры).

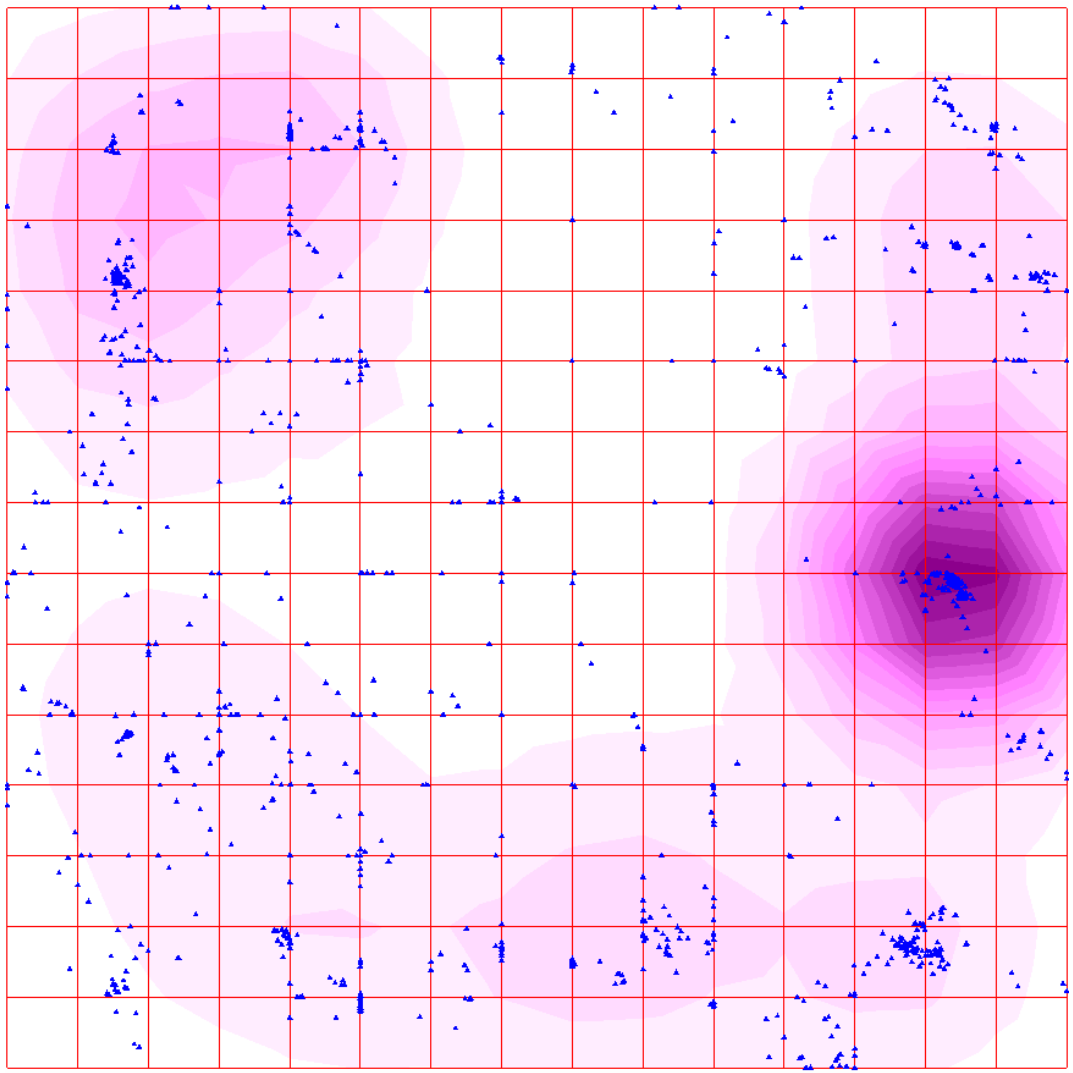


Рис. 1 – Пример упругой карты (наши данные).

Генетический материал

- Для целей нашего исследования использовались гены 16S РНК бактерий. Весь генетический материал был взят из открытой базы данных SILVA (https://www.arb-silva.de/).
- В работе использовались последовательности следующих порядков бактерий: *Acidobacteriales*, *Acidimicrobiales*, *Bacteroidia*, *Chlamydiales*, *Bacillales*.
- Последовательности рРНК, соответствующие бактериям в этих порядках, скачивались из базы данных. Затем с помощью *ad hoc* программы преобразовывались в частотные словари триплетов.

Генетический материал

Индексирование базы данных.

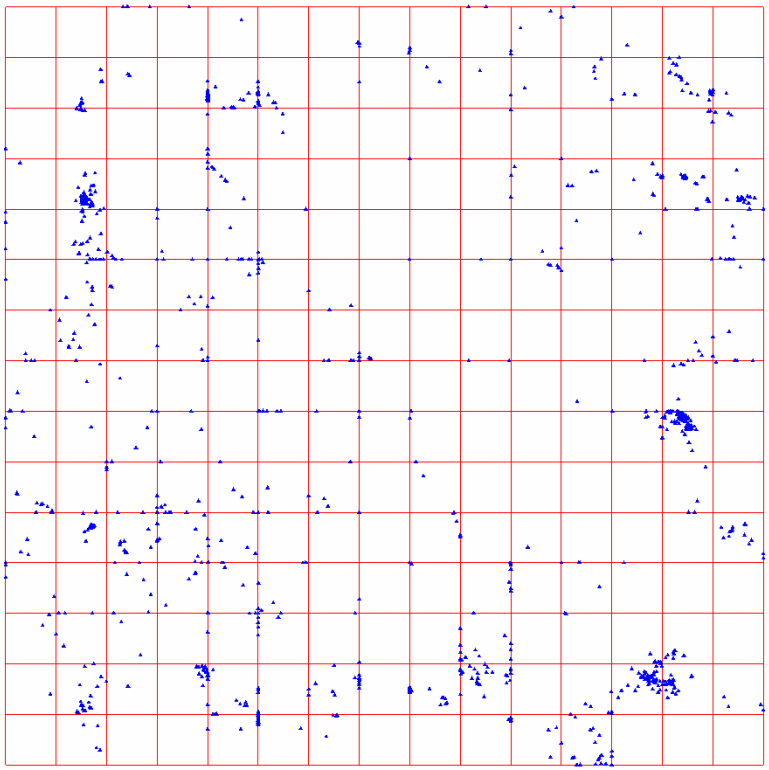
Состав выборки организмов по порядкам до индексирования; N — число организмов.

Порядок	N
Acidobacteriales	34
Acidimicrobiales	24
Bacteroidia	3017
Chlamydiales	820
Bacillales	48579

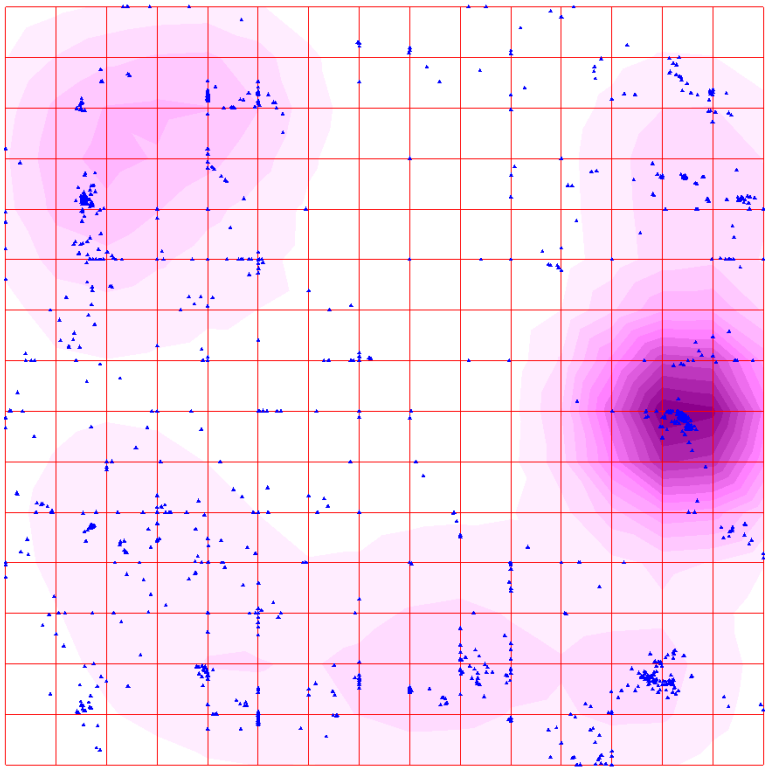
Состав выборки организмов по порядкам после индексирования; N — число организмов.

Порядок	N
Acidobacteriales	34
Acidimicrobiales	24
Bacteroidia	796
Chlamydiales	101
Bacillales	1188

Общая картина



а)



б)

Рис. 2. Распределение всех рассмотренных генов; *а*) — без указания локальной плотности, *б*) — с указанием локальной плотности.

Кластеризация порядков

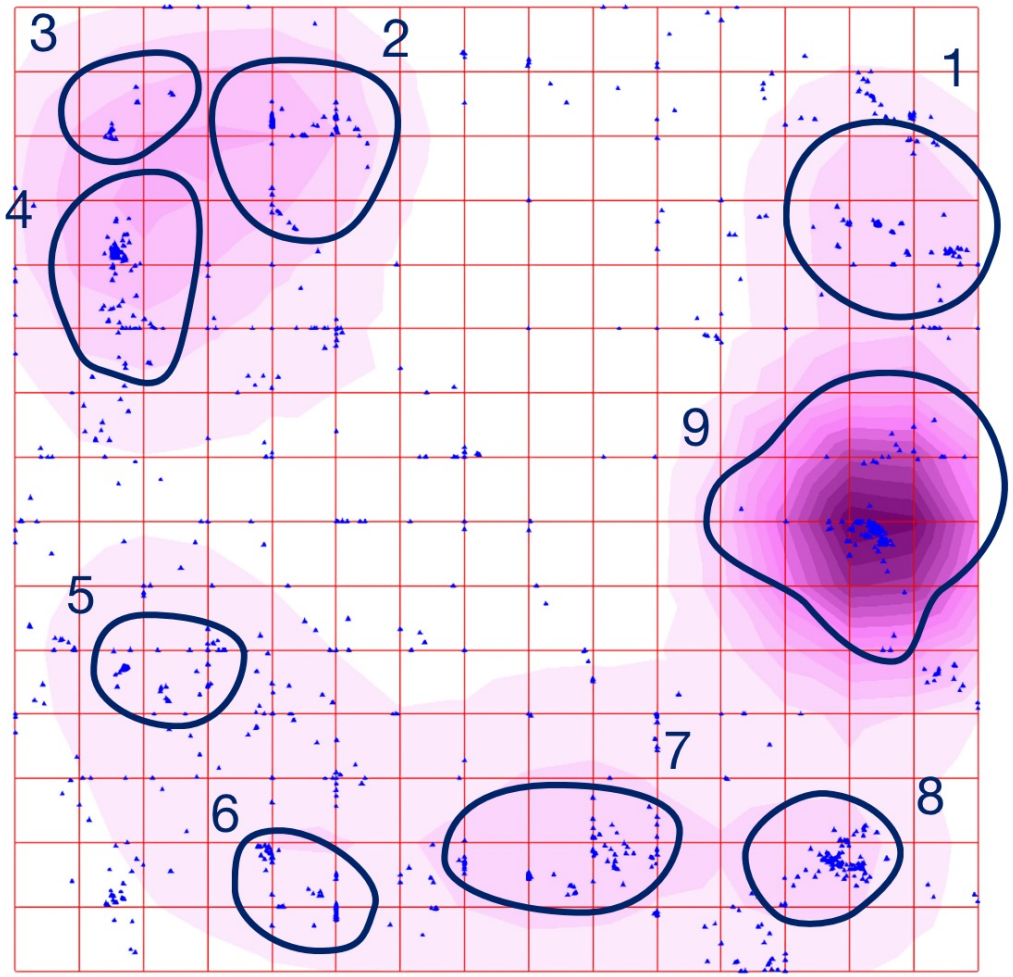


Рис. 3. Кластеры, как мы их видим.

Кластеризация порядков

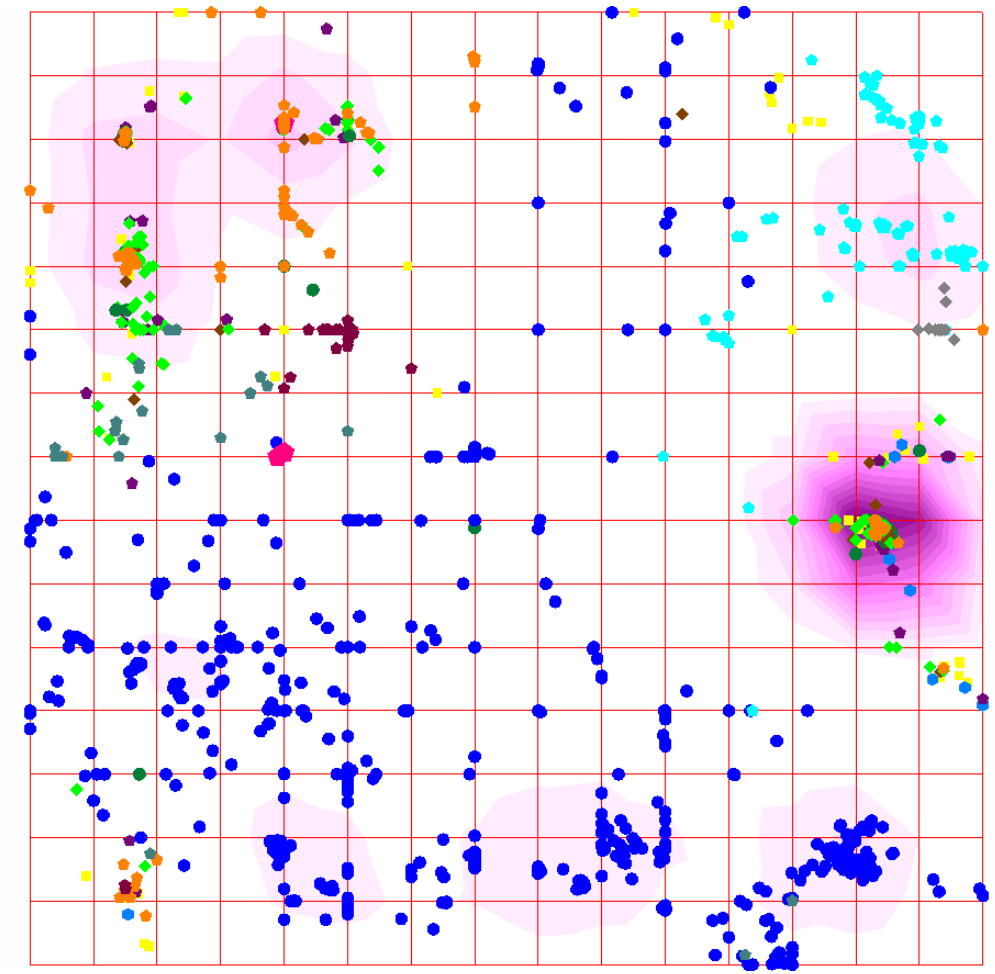
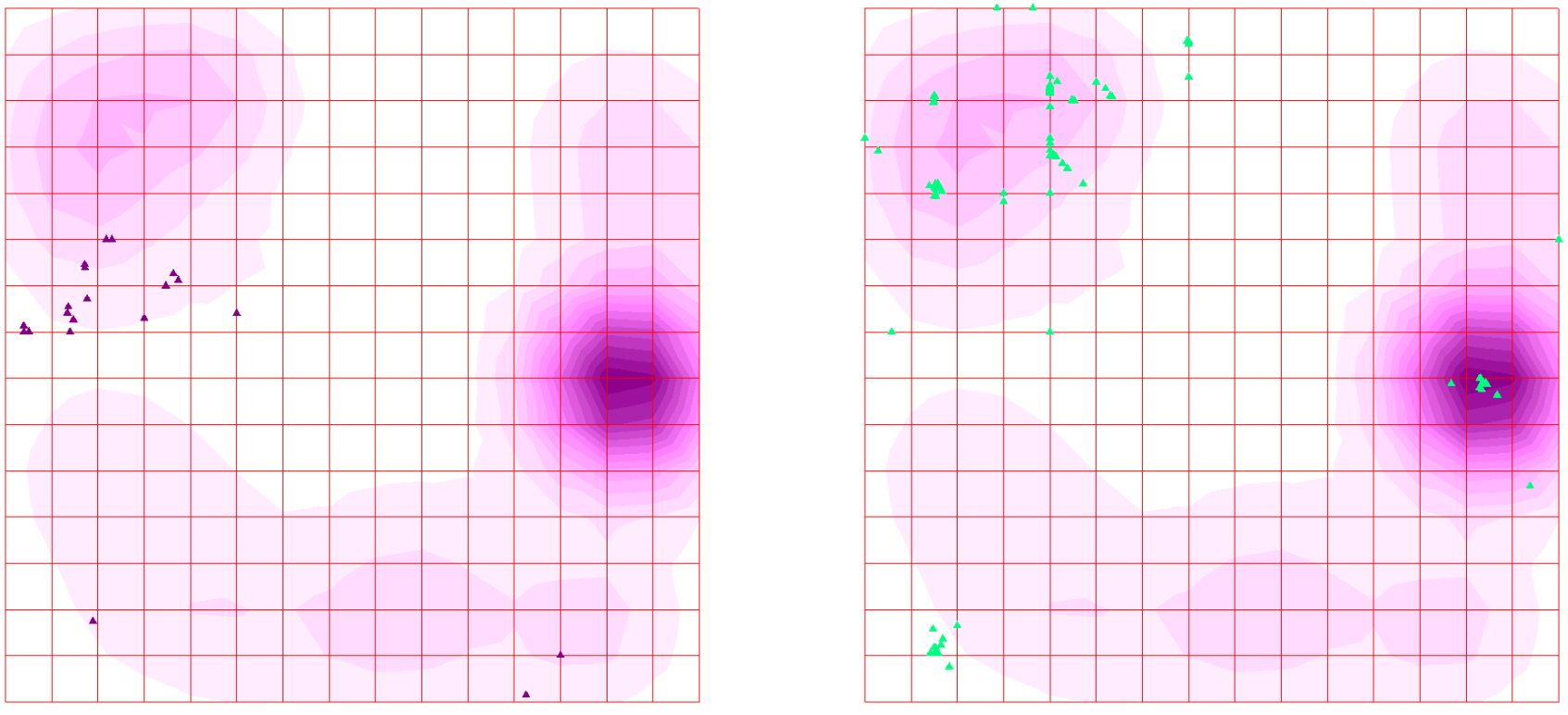


Рис. 4. Кластеры и их таксономическое наполнение.

Кластеризация порядков

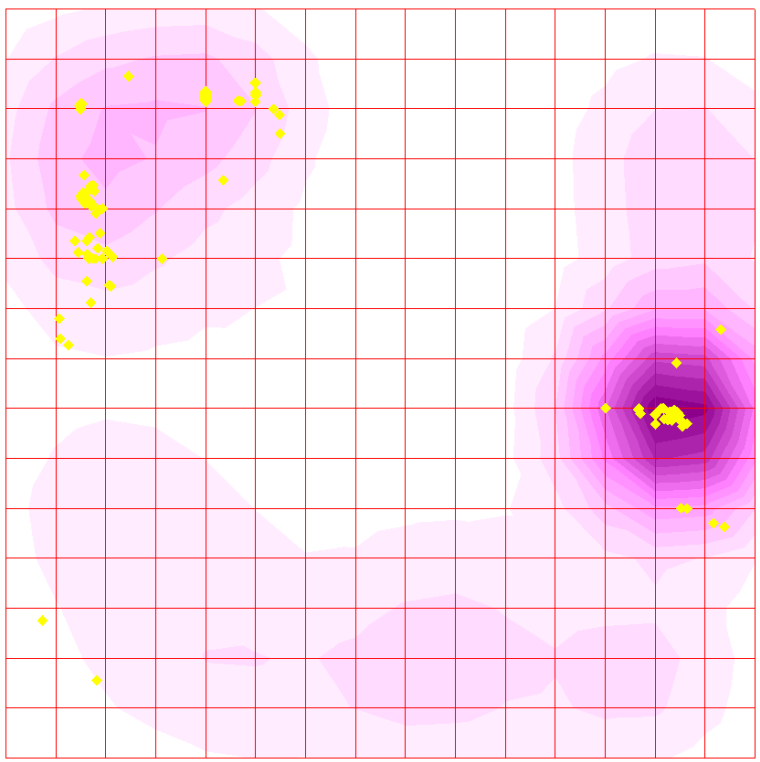


а) *б*)

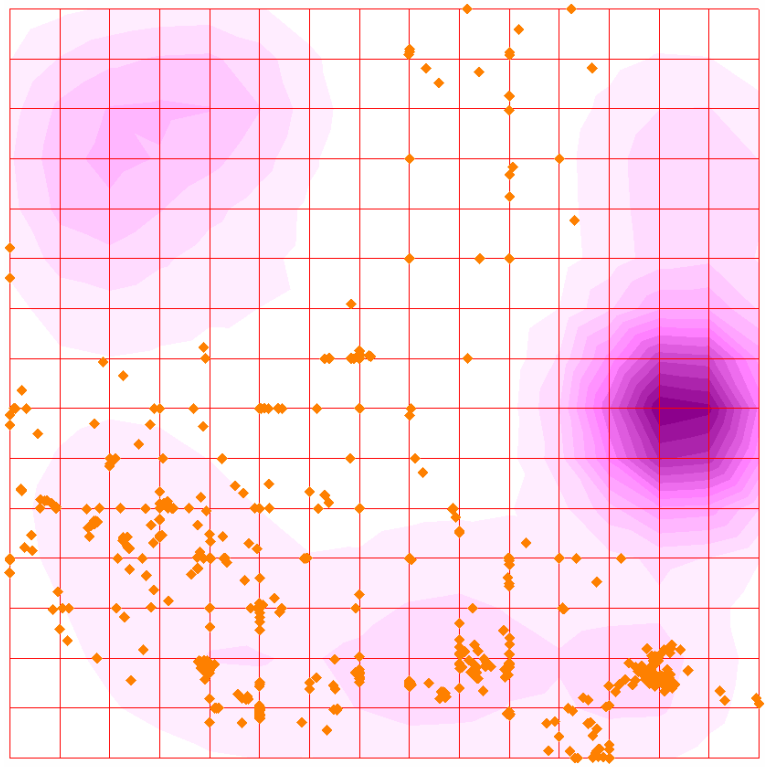
Рис. 5. Распределение порядков; *а*) — *Acidimicrobiales*, *б*) — *Alicyclobacillales*.

.

Кластеризация порядков



в)

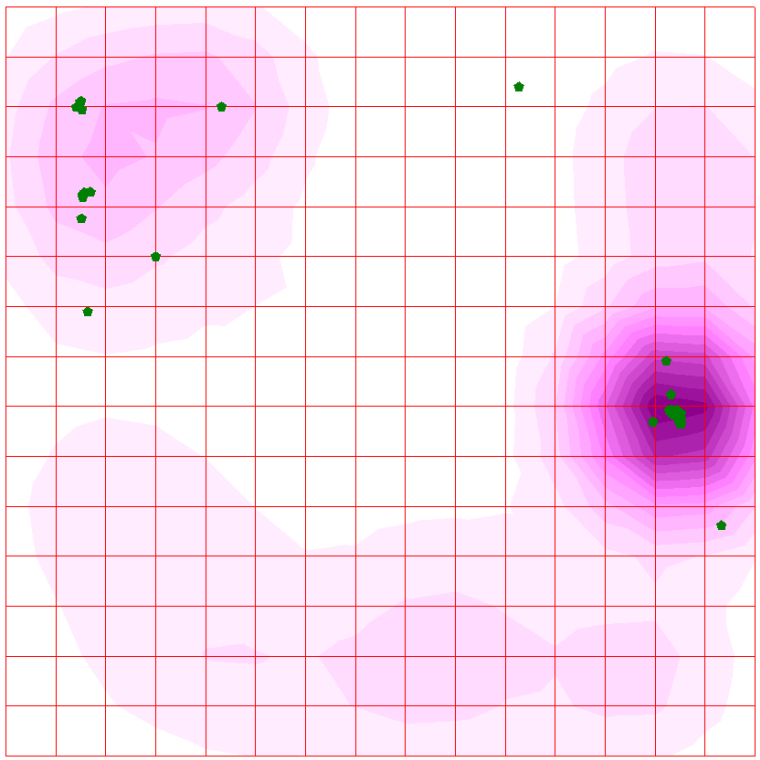


г)

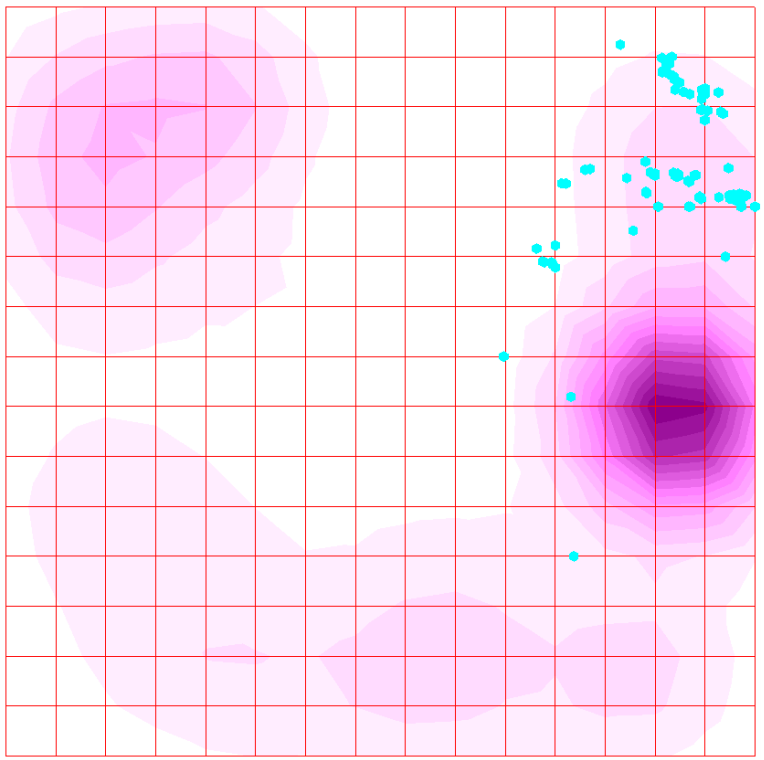
Рис. 5. Распределение порядков; *в*) — *Bacillales*, *г*) — *Bacteroidales*.

.

Кластеризация порядков



д)

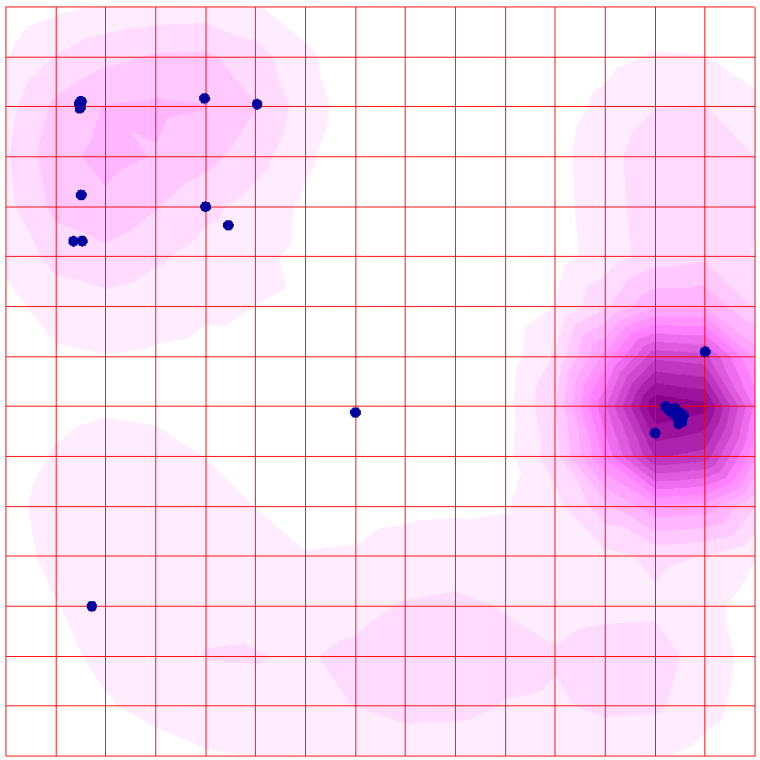


е)

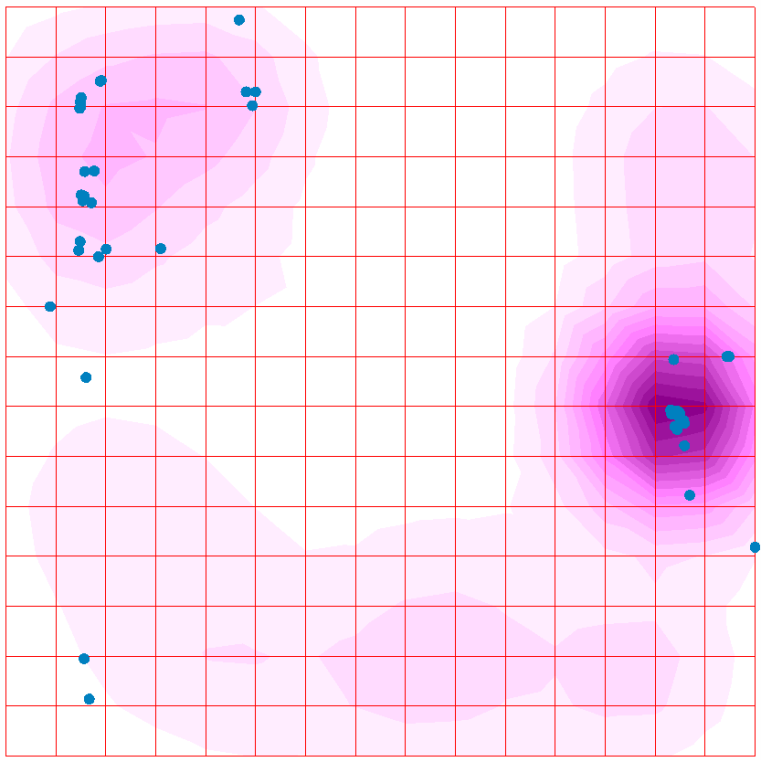
Рис. 5. Распределение порядков; *д*) — *Brevibacillales*, *е*) — *Chlamydiales*.

.

Кластеризация порядков



ж)



з)

Рис. 5. Распределение порядков; *ж*) — *Exiguobacterales*, *з*) — *Lactobacillales*.

.

Кластеризация порядков

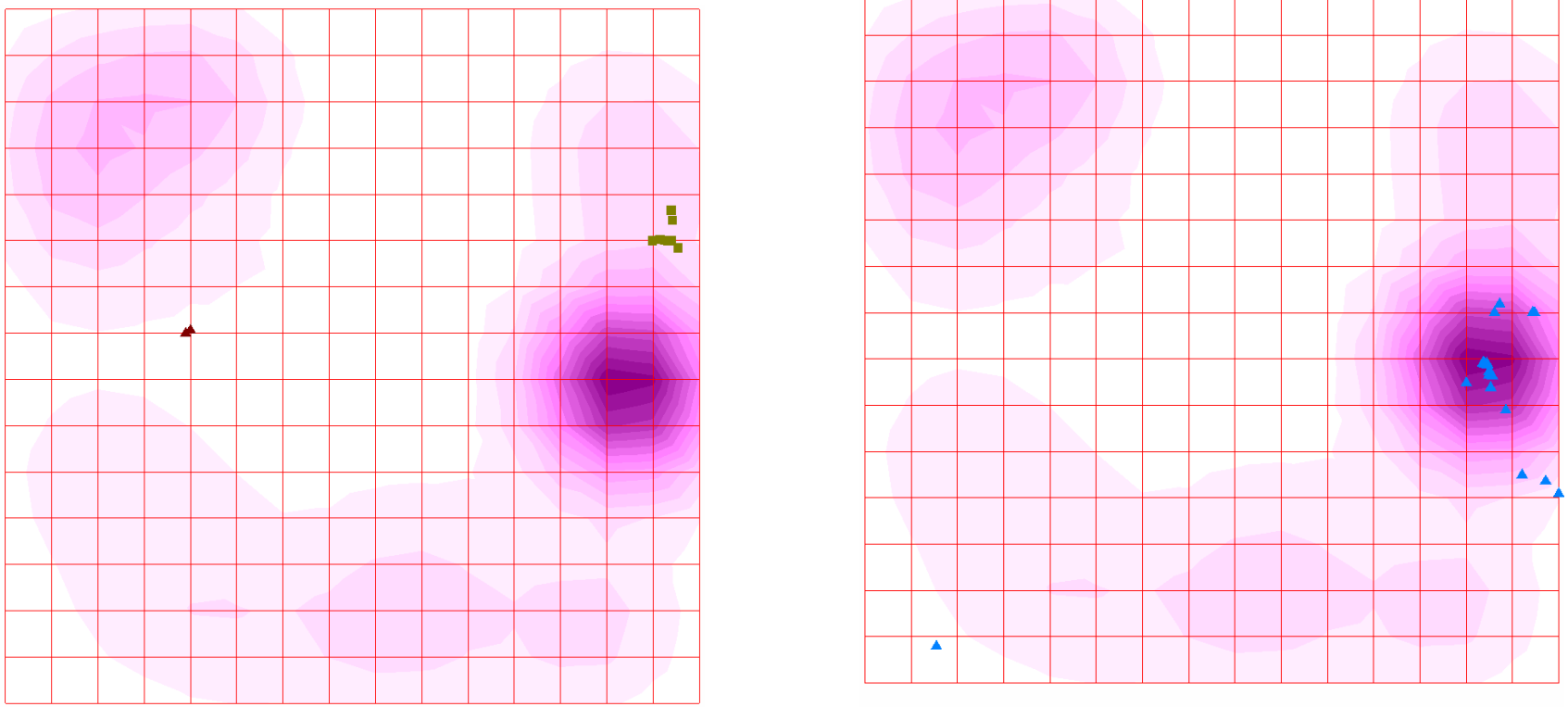
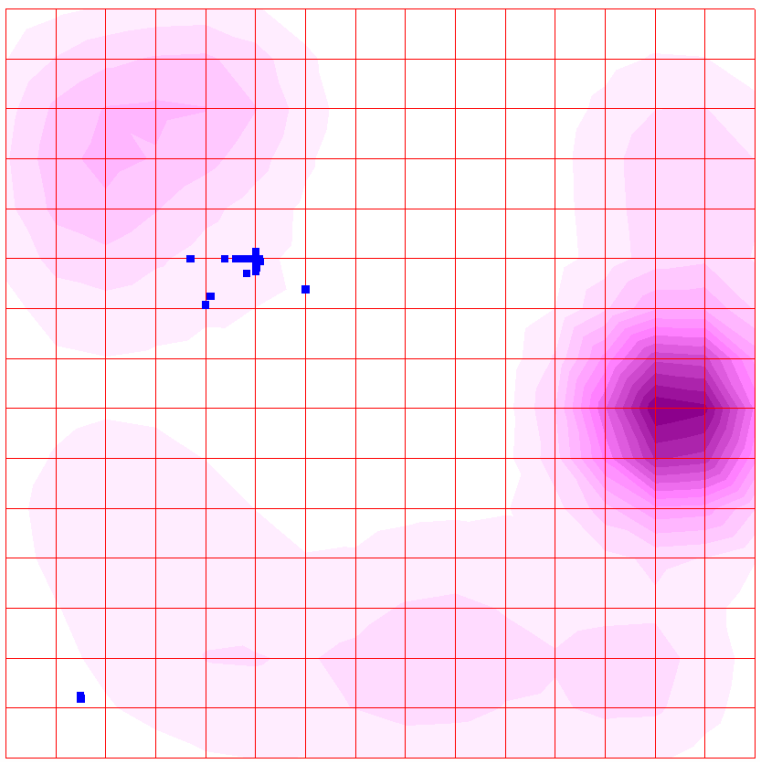


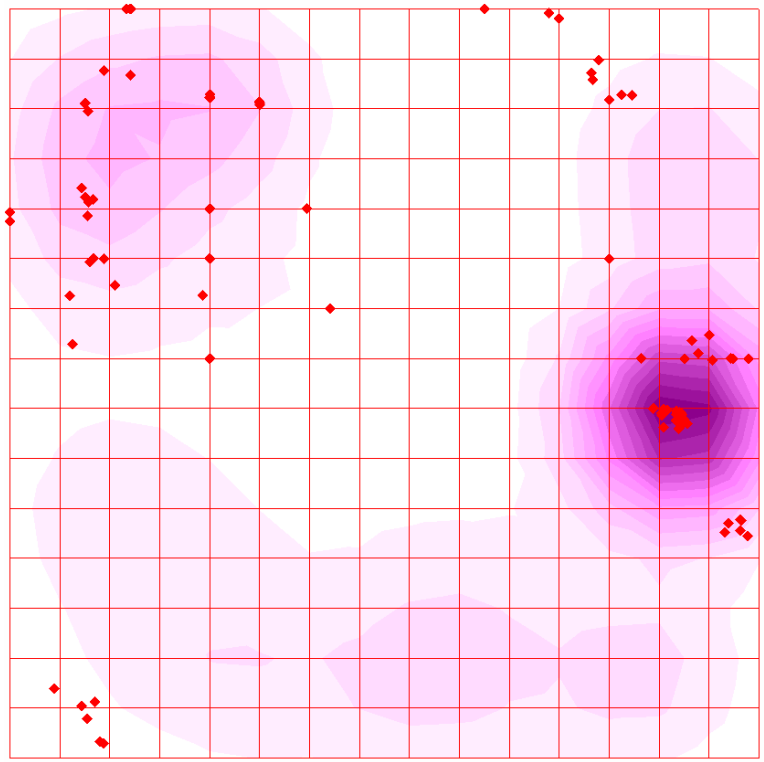
Рис. 5. Распределение порядков;
и) — *Mycoplasmatales* + *Solibacterales*,
к) — *Paenibacillales*.

.

Кластеризация порядков



л)

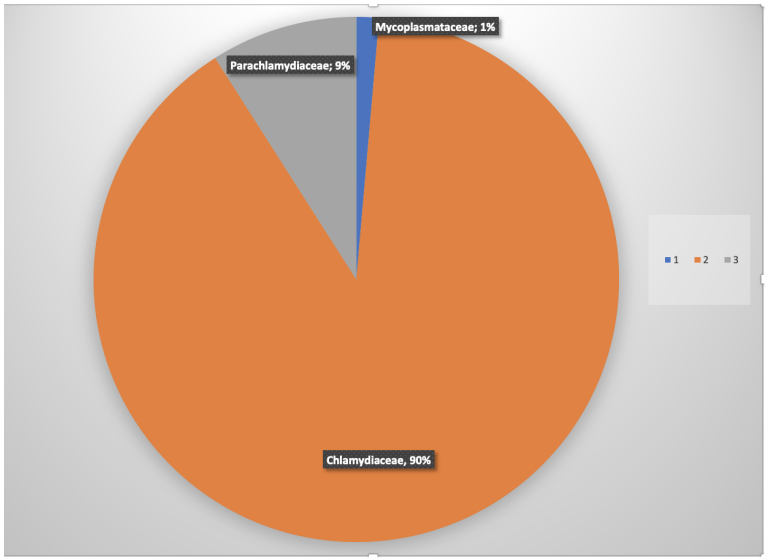


м)

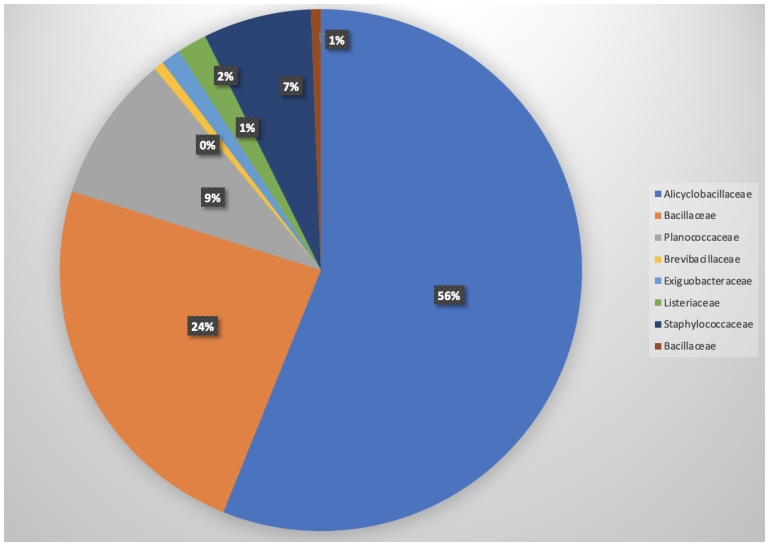
Рис. 5. Распределение порядков; *л*) — *Acidobacteriales*, *м*) — *Staphylococcales*.

.

Диаграммы таксономического состава



Кластер 1



Кластер 2

Рис. 6. Таксономический состав кластеров 1 и 2.

Граф по результатам метода K-средних

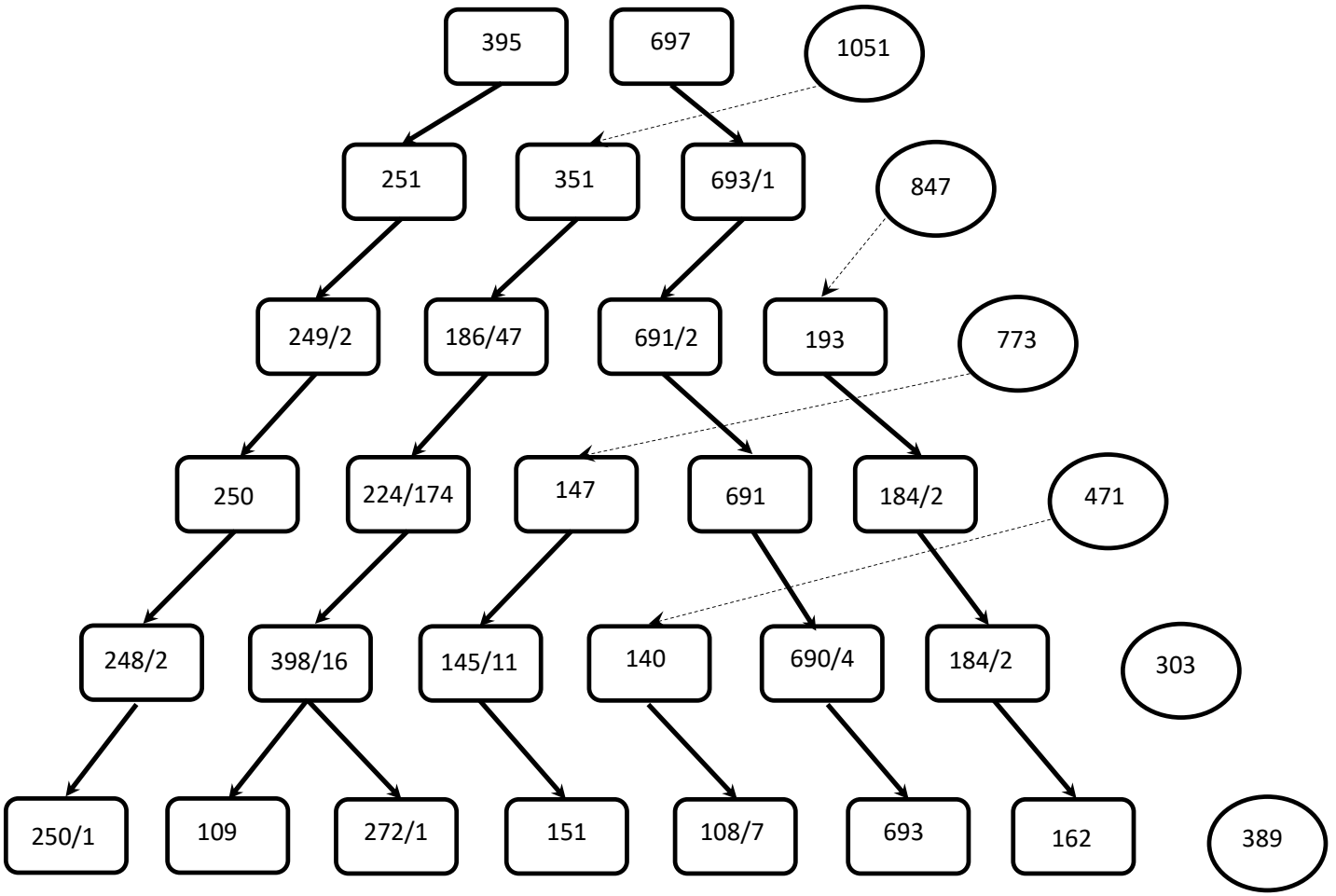


Рис. 7. Структура графа. Коробочки — классы, выделяемые МДЯ (для $2 \leqslant K \leqslant 7$), овалы — волатильные элементы.

Основные результаты для метода упругих карт

- Бóльшая часть последовательностей каждого таксона расходится в разные кластеры;
- Разбиение на кластеры носит неслучайный характер, т. е. в один кластер попадают последовательности одного таксона или близких к нему. Так, например, патогенные бактерии *Chlamydiales* образуют на карте характерный кластер (кластер 1). На карте он выглядит очень обособленно от других кластеров, являясь практически моносоставным (*Chlamydiales* — 99 %, *Bacilli* — 1 %);
- Построенная методом упругих карт кластеризация обладает эффектом масштабирования.

Эффект масштабирования: иллюстрация

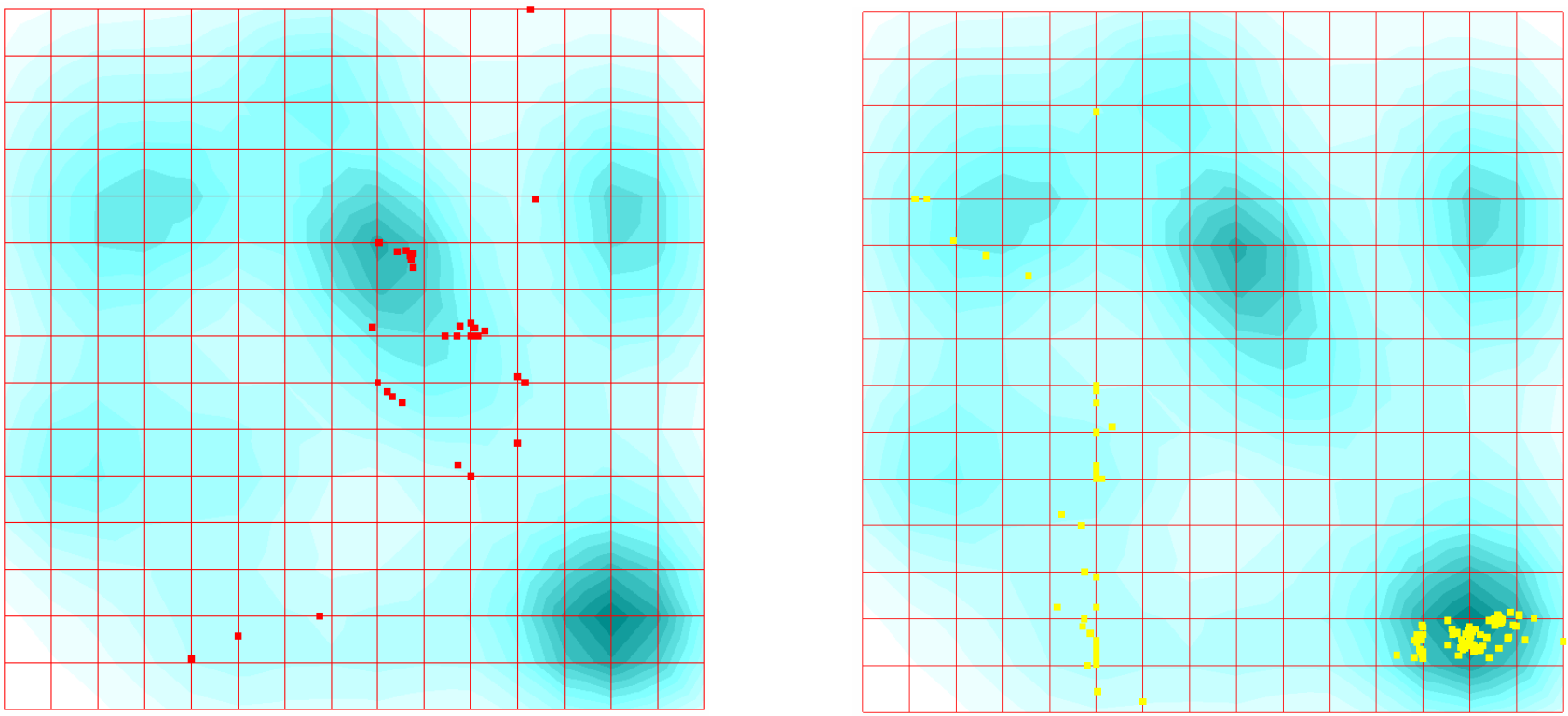


Рис. 8. Распределение двух семейств (*Porphyromonadaceae*, слева и *Bacteroidaceae*, справа) в пределах одного порядка *Bacteroidia*.

Где узнать больше про упругие карты

- A. Gorban, B. Kégl, D. Wünsch, A. Zinovyev (Eds.), Principal Manifolds for Data Visualisation and Dimension Reduction, Lecture Notes in Computational Science and Engineering, Vol. 58, Springer, 2007.
- A. N. Gorban, A. Zinovyev. Principal manifolds and graphs in practice: from molecular biology to dynamical systems // *International Journal of Neural Systems*, Vol. 20, No. 3 (2010) 219 – 232.

Спасибо за внимание

Спасибо за внимание!