Федеральное государственное автономное
образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт фундаментальной биологии и биотехнологии

Кафедра биофизики

УТВЕРЖДАЮ
Заведующий кафедрой
_____ В. А. Кратасюк
«____» _____ 20 __ г.

БАКАЛАВРСКАЯ РАБОТА

06.03.01 Биология

ВЫЯВЛЕНИЕ СВЯЗИ СТРУКТУРЫ ГЕНОМА И ТАКСОНОМИИ ЕГО НОСИТЕЛЯ НА ПРИМЕРЕ ГЕНОМОВ МИТОХОНДРИЙ НЕКОТОРЫХ ВИДОВ ЖИВОТНЫХ

Руководитель _____ д. ф.-м. н., в. н. с. М.Г. Садовский

Выпускник _____ В.С. Федотова

Красноярск 2016

РЕФЕРАТ

Выпускная квалификационная работа по теме «Выявление связи структуры генома и таксономии его носителя на примере геномов митохондрий некоторых видов животных» содержит 40 страниц текстового документа, 22 использованных источника, 11 рисунков, 8 таблиц.

ПОПУЛЯЦИОННАЯ ГЕНОМИКА, ФИЛОГЕНИЯ, КОЭВОЛЮЦИЯ, МЕТОД ДИНАМИЧЕСКИХ ЯДЕР, МЕТОД УПРУГИХ КАРТ.

Цель работы – выявление связи структуры генома и таксономии его носителей на примере геномов митохондрий.

Объектом исследования является связь структуры генома и таксономии его носителей.

Предмет исследования: нуклеотидные последовательности митохондриальных геномов разных видов животных.

Актуальность настоящего исследования обусловлена тем, что ранее ещё не производилось подобных исследований по связи структуры и таксономии на большом генетическом материале.

В итоге была изучена связь и показана синхрония, которая проявлялась в неслучайном распределении таксономических категорий внутри слоистого графа.

СОДЕРЖАНИЕ

Введение

Последние достижения в области современной биологии привели к быстрому росту числа расшифрованных геномов, что, в свою очередь, ставит ряд вопросов перед исследователями в связи с анализом обширной генетической информации. Этот извлечение знаний из генетических систем и упорядочивание больших массивов данных.

Актуальность работы обусловлена как появлением большого числа вновь расшифрованных геномов, которые позволяют ставить задачи популяционного и эволюционного анализа, так и необходимостью развития новых методов анализа данных, использующих информацию, содержащуюся во всём геноме в целом, а не в отдельных его частях.

Объектом работы является связь структуры генома органеллы и таксономии его носителя. Предметом исследования являются нуклеотидные последовательности митохондриальных геномов разных видов животных и те виды связей, которые выявляются на множестве геномов и между геномами и таксономией их носителя.

Цель настоящей работы — выявление, описание и анализ связи структуры генома органеллы и таксономии его носителей на примере геномов митохондрий.

Для достижения указанной цели были решены следующие задачи:

- создание базы геномов и их частотных словарей;
- создание репрезентативной выборки с равномерным распределением видов в родах для кластеризации;
- кластеризация митохондриальных геномов методом динамических ядер (*k*-means) от 2 до 8 классов;
- построение упругой карты таксономического типа *Chordata* и изучение её особенностей;
- построение слоистого графа кластеризации и изучение его свойств.

По материалам бакалаврской работы опубликована 1 статья, основные результаты работы докладывались на следующих конференциях:

- седьмая международная школа молодых учёных «Системная биология и биоинформатика» SBB’2015;
- международная конференция студентов, аспирантов и молодых учёных «Проспект Свободный — 2015»;
- третья международная конференция IWBBIO 2015.

Основная часть

1. Обзор литературы

1.1. Митохондрии

Митохондрии — энергетическая система клетки, органеллы, синтезирующие АТФ. Их основная функция связана с окислением органических соединений энергии для синтеза молекул АТФ. Исходя из этого, митохондрии часто называют энергетическими станциями клетки, а также органеллами клеточного дыхания.

Основной функцией митохондрий является синтез АТФ, который образуется в результате процессов окисления органических субстратов и фосфорилирования АДФ. Системы дальнейшего переноса электронов и сопряжённого с ним фосфорилирования АДФ располагаются в мембранах крист митохондрий.

В матриксе митохондрий локализуется автономная система митохондриального белкового синтеза. Она представлена молекулами ДНК, свободными от гистонов. Это сближает их с ДНК бактериальных клеток. Подавляющее большинство белков митохондрий находится под генетическим контролем клеточного ядра и синтезируется в цитоплазме. Митохондриальная ДНК кодирует лишь 13 митохондриальных белков, которые локализованы в мембранах и представляют собой структурные белки (интеграция белковых комплексов в мембрану) [1].

1.2. Ядерный и митохондриальный геномы

Способность передавать признаки по наследству является естественным свойством любого живого организма. Наследственная информация, необходимая для создания нового организма содержится в нуклеиновых кислотах. Их совокупность называют геномом. Функционально геном состоит из генов. Каждый ген представлен последовательностью

нуклеотидов некоторого участка нуклеиновой кислоты и кодирует информацию об одном белке. Отдельная молекула нуклеиновой кислоты из общего разнообразия, составляющего геном, может содержать значительное количество генов. Размеры ядерных геномов живых организмов варьируют в широком диапазоне. Так, например, геном бактерии *Micoplasma* состоит всего из 500 генов, а геном человека из более чем 25 000 [2].

Митохондриальная ДНК в большинстве случаев представлена кольцевыми молекулами. Лишь у немногих видов, в частности некоторых кишечнополостных животных, эти молекулы линейные. У животных размеры молекул митохондриальной ДНК варьируют незначительно, обычная величина — около 16 тысяч пар нуклеотидов.

В митохондриальной ДНК млекопитающих и других животных всего 37 генов: 13 из них кодируют субъединицы белков – ферментов окислительного фосфорилирования, 2 гена кодируют рибосомные РНК и 22 небольших гена — транспортные РНК. Такой же набор генов присутствует в митохондриальной ДНК высших растений, к нему добавляется еще ген 5 S РНК. По размеру молекул митохондриальная ДНК растений значительно больше, чем митохондриальная ДНК животных: от 200 тысяч пар нуклеотидов у видов рода *Brassica* до 2500 тысяч пар нуклеотидов у арбуза. Увеличение размера молекул митохондриальных ДНК происходит за счет некодирующих последовательностей, кроме них в митохондриальной ДНК растений включены фрагменты хлоропластной ДНК [3].

1.3. Сравнительная характеристика филогении и таксономии

Филогения — это раздел биологии, изучающий родственные взаимоотношения разных групп живых организмов. Филогению обычно отображают в виде «эволюционных деревьев» или систематических названий.

Все методы построения филогенетических древ базируются на молекулярных последовательностях белков или кодирующих участках ДНК

(например, PTMS — phylogenetic tree reconstruction from molecular sequences) [4]. Существует несколько основных алгоритмов построения филогенетических деревьев:

1) методы, основанные на оценке расстояний (матричные методы);
2) наибольшего правдоподобия (maximal likelihood, ML) [5];
3) максимальной экономии (maximal parsimony, MP) [6].

В первой группе методов вычисляются эволюционные расстояния между всеми листьями и строится дерево, в котором расстояния между вершинами наилучшим образом соответствует матрице попарных расстояний. Во второй группе методов используется модель эволюции и строится дерево, которое наиболее правдоподобно при данной модели. В третьей группе методов выбирается дерево с минимальным количеством мутаций, необходимых для объяснения данных.

Все вышеперечисленные методы могут быть достаточно точными, но в любом случае они исключают некодирующие участки ДНК. Также подбор последовательностей геномов довольно специфичен: подбираются близкие последовательности, в количестве не большем 50. Для простоты чаще работают с белками.

Нашей задачей была кластеризация частот полных геномных последовательностей, поэтому нам необходимо было использовать такие методы, которые не опираются на методы выравнивания геномов. В первую очередь необходимо полное извлечение информации из статистических данных, т.е. решение биологических задач с помощью чисто математических подходов. Необходимо выделение структурированности данных, для этого используются методы кластерного анализа. Кластеризация методом динамических ядер позволяет работать как с частотными словарями генов, так и с частотными словарями полных нуклеотидных последовательностей.

Таксономия — это наука о категоризации и классификации вещей (в нашем случае животных), основанных на заранее определённых системах.

Таксономия базируется на морфологических, поведенческих, генетических и биохимических особенностях организмов.

Методы филогении позволяют составлять новую систематику животных, увеличение точности этих методов позволит представить полную картину родственных связей между видами. Тем самым филогению можно рассматривать как мягкий метод построения таксономии. В данном исследовании проводится построение филогении, которое в свою очередь опирается на базовую таксономию организмов.

В банках данных существует огромное количество информации, большая часть которой относится либо к каким-то участкам ядерной ДНК, либо к ДНК органелл (митохондрии, хлоропласты). На данный момент количество полностью отсеквенированных ядерных геномов сравнительно невелико, что не позволяет пока применить метод по отношению к ним.

Метод чувствует разницу между довольно низкими таксонами (семейства, рода) на уровне триплетов. Структурное распределение видов показывало, что близкие между собой геномы определялись в одну группу.

1.4. Методы анализа многомерных данных

За последнее время анализ многомерных данных стал одним из основных направлений прикладной математики, которое активно развивается и применяется практически во всех областях исследований. Анализ многомерных данных (или MDA — Multivariate Data Analysis) является одной из наиболее популярных и востребованных междисциплинарных областей знания и активным инструментом синтеза различных дисциплин [7].

Одним из наиболее известных методов анализа многомерных данных является метод главных компонент и его обобщения для нелинейных случаев. Методы анализа многомерных данных реализуются в тесной взаимосвязи и взаимодействии с методами факторного и кластерного анализа.

Цель анализа данных — извлечение содержащейся информации. Задача снижения размерности набора данных — описание каждой точки данных с помощью величин, число которых меньше размерности пространства и которые являются функциями исходных координат. Метод главных компонент позволяет уменьшить размерность данных, потеряв при этом наименьшее количество информации. Вычисление главных компонент может быть проведено с помощью нескольких алгоритмов (например, итерационный алгоритм сингулярного разложения и сингулярное разложение тензоров) [8].

Метод применим всегда, однако при заданных ограничениях на точность не всегда эффективно снижает размерность. Прямые и плоскости не всегда обеспечивают хорошую аппроксимацию данных.

Основной задачей данного исследования являлась кластеризация данных. Главным методом для решения задачи был выбран метод динамических ядер, который относится к процедурам прямой классификации. Для визуализации данных и решения обратной задачи был выбран метод упругих карт.

1.5. Приёмы выделения структурированности данных

Для выделения структурированности большого объёма данных применяются методы кластерного анализа. Это статистическая процедура, выполняющая сбор данных с информацией о выборке объектов и затем упорядочивающая эти объекты в сравнительно однородные группы.

В настоящее время существует огромное количество алгоритмов кластер-анализа. Кластер представляет собой некоторую целостность, причём разные кластеры могут касаться друг друга и даже пересекаться [9]. Свойство кластеризации заключается в том, что точки одного кластера находятся на относительно близком расстоянии друг от друга, в то время как точки из разных кластеров — на большом.

Все алгоритмы кластеризации можно поделить на две группы, которые основаны на двух принципиально разных стратегиях.

Первая группа— это иерархические или агломерационные алгоритмы. Кластеры объединяются на основе их «близости». В процессе постепенно ослабляется критерий о том, какие объекты являются уникальными, а какие нет. Другими словами, понижают порог, относящийся к решению об объединении двух или более объектов в один кластер. В результате, связываются вместе всё большее и большее число объектов и агрегируется все больше и больше кластеров, состоящих из все сильнее различающихся элементов. Процесс объединения останавливается, когда дальнейшее объединение приводит к нежелательным по некоторым причинам кластерам. Например, можно остановиться, когда имеется уже определённое количество кластеров, или же можно использовать меру компактности для кластеров и отказаться от построения кластера путём объединения двух небольших кластеров, если в результате кластер имеет точки, которые разбросаны по слишком большой области.

Ко второй группе относятся алгоритмы, в которых точки рассматриваются в определённом порядке и каждой из них назначается наиболее подходящий кластер. Этому процессу предшествует короткий этап определения начальных кластеров. Вариации допускают случайное объединение или разделение кластеров, если они находятся на периферии.

Также алгоритмы кластеризации можно выделить иным способом. Объединение или метод древовидной кластеризации используется при формировании кластеров несходства или расстояния между объектами. Эти расстояния могут определяться в одномерном или многомерном пространстве. Наиболее прямой путь вычисления расстояний между объектами в многомерном пространстве состоит в вычислении евклидовых расстояний. Если имеется двух- или трёхмерное пространство, то эта мера является реальным геометрическим расстоянием между объектами в пространстве. Наиболее используемый тип расстояния — Евклидово

расстояние. Оно попросту является геометрическим расстоянием (d) в многомерном пространстве и для точек $x = (x_1, \dots, x_n)$, $y = (y_1, \dots, y_n)$ вычисляется следующим образом:

$$d(x, y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}. \qquad (1)$$

Евклидово расстояние вычисляется по исходным, а не по стандартизованным данным. Этот способ вычисления имеет определенные преимущества (например, расстояние между двумя объектами не изменяется при введении нового объекта, который может оказаться «лишним»).

Одним из наиболее известных методов кластеризации является метод динамических ядер. Метод подразумевает под собой уже известное число кластеров и использует евклидову метрику. В общем случае метод строит кластеры на возможно больших расстояниях друг от друга. С вычислительной точки зрения этот метод начинает с K случайно выбранных кластеров, а затем изменяет принадлежность объектов к ним, чтобы минимизировать изменчивость внутри кластеров и максимизировать изменчивость между кластерами. Метод динамических ядер хорошо подходит для работы с однородными многомерными данными. Для проверки точности результатов определяется различимость классов, которая в данной работа подсчитывалась в разделе 2.4 [10].

2. Материалы и методы

2.1. Генетический материал

Используемый в работе генетический материал был взят из EMBL банка в количестве 3726 (по состоянию на 27.06.2014) митохондриальных геномов. Для кластеризации было отобрано 2990 геномов. Ниже (раздел 3.1) будет описываться метод отбора геномов для кластеризации.

2.2. Частотные словари

В исследованиях эволюционных процессов и их молекулярного уровня интригующей является проблема связи между структурой нуклеотидной последовательности того или иного организма и его таксономическим положением, которое определяется классическими методами по различным морфологическим признакам.

При изучении связи между структурой нуклеотидной последовательности и функцией, определяемой этой последовательностью, функция, как правило, понимается исследователями одинаково. Также не возникает разночтений в понимании исследователями того, что есть таксономическое положение носителя нуклеотидной последовательности.

Изучение структуры нуклеотидной последовательности требует определения каких-либо дополнительных отношений (например, классификаций) на множестве таких последовательностей. Уже раннее проведённые исследования показали, что большая совокупность нуклеотидных последовательностей может быть разбита на несколько классов, в каждом из которых эти последовательности близки друг к другу в смысле близости их реальных частотных словарей в евклидовой метрике.

Каждая нуклеотидная последовательность является символьной последовательностью из четырёхбуквенного алфавита $\aleph = \{A, C, G, T\}$ той же длины N — генетический текст. Любую связную подпоследовательность длины q из рассматриваемой последовательности будем называть словом или

q-плетом. Сопоставим каждое слово и его частоту (число его копий в изучаемом тексте, отнесённое к общему числу разных слов в данном тексте); такой список всех слов длины q, входящих в данную последовательность, вместе с частотами их встречаемости назовём частотным словарём толщины q и назовём W_q [11 – 13]. Всюду далее мы будем изучать словари W_3 триплетов. В настоящей работе частотные словари составлялись с помощью скрипта, найденного на открытых ресурсах программного обеспечения по биоинформатике.

2.3. Индексированная база данных

Исходная база митохондриальных геномов содержит разное количество видов в различных таксономических родах: например, в роде *Equus* содержится 257 геномов, а в роду *Ablennes* — всего 1 геном. Изучалась устойчивость классификации методом динамических ядер в зависимости от полноты представленности разных родов. Для этого в работе использовались две базы данных: исходная, полученная из банка, и отредактированная, которая содержала не больше 50 видов в одном роде. Это ограничение определялось по структуре исходной базы геномов. Для ограничения было взято число видов, на порядок превышающее наиболее часто встречающееся число видов в родах. Число геномов в отредактированной базе — 2990.

Таблица 1 – Распределение геномов в таксономических семействах в исходной базе

Среднее арифметическое	Минимум	Максимум
6,54	1	287

В таблице 1 указано среднее число видов в таксономических семействах, минимум и максимум числа видов в исходной базе данных. В таблице 2 также указано среднее число видов, минимум и максимум видов в таксономических семействах, но уже индексированной базы данных.

Таблица 2 – Распределение геномов в таксономических семействах в индексированной базе

Среднее арифметическое	Минимум	Максимум
5,22	1	195

В индексированной базе данных значительно уменьшилось максимальное число видов в семействах, среднее арифметическое снизилось на единицу.

Отбор геномов проводился для получения более точных результатов. В случае кластеризации исходной базы данных многочисленные геномы одного вида в 63-мерном пространстве образовывали отдельный кластер, который притягивал к себе окружающие его точки и вносил существенные искажения в картину распределения геномов.

2.4. Метод динамических ядер

Одним из самых наиболее хорошо известных методов кластеризации является метод динамических ядер. Этот метод был также одним из самых часто использующихся с 1950-х годов. Идея метода была предложена Штейнгаузом в 1956 году, но алгоритм, который так часто используется сегодня, не публиковался вплоть до 1982 года (публикация Ллойда) [14].

Построение отношений на множестве объектов — одна из актуальных областей применения искусственного интеллекта и методов анализа многомерных данных. Первым и наиболее распространенным примером этой задачи является классификация без учителя. Задан набор объектов, каждому объекту сопоставлен вектор значений признаков (строка таблицы). Требуется разбить эти объекты на классы эквивалентности. Следует подчеркнуть, что и выбор метода, и интерпретация полученных результатов будут зависеть от целей и задач разбиения и того, что мы будем делать с его результатом. Ответ на этот вопрос позволяет приступить к формальной постановке задачи, которая всегда требует компромисса между сложностью решения и точностью формализации: буквальное следование содержательному смыслу

задачи нередко порождает сложную вычислительную проблему, а следование за простыми и элегантными алгоритмами может привести к противоречию со здравым смыслом. Для каждого нового объекта мы должны сделать два дела: 1) найти класс, к которому он принадлежит; 2) использовать новую информацию, полученную об этом объекте, для исправления (коррекции) правил классификации.

Пусть $\{x^p\}$ — векторы значений признаков для рассматриваемых объектов и в пространстве таких векторов определена мера их близости $\rho\{x, y\}$. Для определенности примем, что чем ближе объекты, тем меньше ρ. С каждым классом будем связывать его типичный объект. Далее будем называть его ядром класса. Требуется определить набор из m ядер $y_1, y_2, \dots y_m$ и разбиение $\{x^p\}$ на классы: $\{x^p\} = Y_1 \cup Y_2 \cup \dots \cup Y_m$, минимизирующее следующий критерий

$$Q = \sum_{i=1}^{m} D_i \to min, \qquad (2)$$

где для каждого (i-го) класса D_i — сумма расстояний от принадлежащих ему точек выборки до ядра класса:

$$D_i = \sum_{x^p \in Y_i} p\left(x^p, y^i\right). \qquad (3)$$

Минимум Q берется по всем возможным положениям ядер y_iи всем разбиениям $\{x_p\}$на m классов Y_i. Если число классов заранее не определено, то полезен критерий слияния классов: классы Y_i и Y_j сливаются, если их ядра ближе, чем среднее расстояние от элемента класса до ядра в одном из них.

Сетевые алгоритмы классификации без учителя строятся на основе итерационного метода динамических ядер. Пусть задана выборка предобработанных векторов данных $\{x_p\}$. Пространство векторов данных обозначим E. Каждому классу будет соответствовать некоторое ядро a. Пространство ядер будем обозначать A. Для каждых $x \in E$ и $a \in A$ определяется мера близости $d(x, a)$. Для каждого набора из k ядер $a_1, \dots, a_k$ и любого разбиения $\{x_p\}$ на k классов $\{x_p\} = P_1 \cup P_2 \cup \dots \cup P_k$ определим критерий качества [15]:

$$D = D(a_1, a_2, \dots, a_k, P_1, P_2, \dots P_k) = \sum_{i=1}^{k} \sum_{x \in P_i} d(x, a_i) \qquad (4)$$

Задача поиска ядра для данного класса P имеет своим решением

$$a_p = {\sum_{x \in P} x}/{\left\| \sum_{x \in P} x \right\|} \qquad (5)$$

Известно, что в общем случае метод динамических ядер не даёт единственности построения классификации: результат зависит от начального распределения геномов по классам (которое каждый раз определяется случайным образом). Другой проблемой является определение минимального числа классов, на которое следует разделить геномы [16]. В рамках настоящей работы не проводилось никаких специальных исследований, направленных на выявление оптимального числа классов, получаемых методом динамических ядер. Количество классов бралось произвольным образом равным восьми.

Таблица 3 – Проверка различимости классов кластеризации, 2 класса

Данные измерений \ Виды	Все виды	Устойчивые виды
Расстояние между классами	0,093	0,094
Радиус класса 1	0,132	0,124
Радиус класса 2	0,128	0,113
Сумма радиусов	0,26	0,236

В работе проверялась различимость построенных при кластеризации классов как для полной базы, так и исключительно для устойчиво делящихся видов. Введём понятие различимости классов. Классы считались различимыми, если их радиусы не больше расстояния между их центрами. В связи с этим классы хорошо различимы в случае, когда сумма их радиусов не больше расстояния между их центрами. В таблице 3 предоставлены измерения радиусов и расстояния между классами. В таблицах 4 и 5 также отображены измерения радиусов окружностей и расстояний между классами.

Таблица 4 – Проверка различимости классов кластеризации, 3 класса кластеризации

Данные измерений \ Виды	Все виды	Устойчивые виды
Расстояние между классами 1-2	0,106	0,108
Расстояние между классами 1-3	0,104	0,105
Расстояние между классами 2-3	0,097	0,098
Радиус класса 1	0,133	0,117
Радиус класса 2	0,05	0,047
Радиус класса 3	0,072	0,069
Сумма радиусов 1-2	0,184	0,165
Сумма радиусов 1-3	0,2	0,186
Сумма радиусов 2-3	0,122	0,116

Таблица 5 – Проверка различимости классов кластеризации, 4 класса

Данные измерений \ Виды	Все виды	Устойчивые виды
Расстояние между классами 1-2	0,106	0,107
Расстояние между классами 1-3	0,081	0,084
Расстояние между классами 1-4	0,15	0,15
Расстояние между классами 2-3	0,082	0,086
Расстояние между классами 2-4	0,131	0,132
Расстояние между классами 3-4	0,078	0,075
Радиус класса 1	0,134	0,118
Радиус класса 2	0,067	0,064
Радиус класса 3	0,031	0,025
Радиус класса 4	0,017	0,016
Сумма радиусов 1-2	0,202	0,182
Сумма радиусов 1-3	0,165	0,143
Сумма радиусов 1-4	0,15	0,134
Сумма радиусов 2-3	0,01	0,09

Сумма радиусов 2-4	0,084	0,081
Сумма радиусов 3-4	0,048	0,042

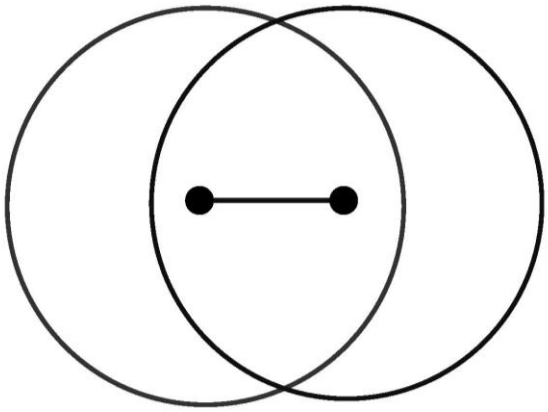


Рисунок 1 – Различимость 2-х классов кластеризации всех видов

На рисунке 1 изображена различимость 2-х классов кластеризации всех видов, окружности имеют область перекрытия. Если посмотреть на площадь этой зоны, то можно предположить, что там находятся волатильные геномы.

В целом получилась средняя различимость для всех этапов, у всех есть зоны перекрытия. Различимость устойчивых видов более явная, если сравнивать с различимостью классов всех видов.

2.5. ViDaExpert

Задачи выполнялись с помощью программы ViDaExpert. Разработка программного обеспечения ViDaExpert представляет собой непрерывный процесс, который был начат в 2000 году Андреем Зиновьевым для его кандидатской диссертации в ИВМ СО РАН. На данный момент проект активно используется исследователями института Кюри для анализа данных [17]. Данная программа была использована для построения упругой карты и кластеризации методом динамических ядер.

Упругая карта служит для нелинейного сокращения размерности данных. В многомерном пространстве данных располагается поверхность,

которая приближает имеющиеся точки данных и при этом является, по возможности, не слишком изогнутой. Данные проецируются на эту поверхность и потом могут отображаться на ней, как на карте. Ее можно представлять себе как упругую пластину, погруженную в пространство данных и прикрепленную к точкам данных пружинками. Она служит обобщением метода главных компонент (в котором вместо упругой пластины используется абсолютно жесткая плоскость).

По построению, упругая карта представляет собой систему упругих пружин, вложенную в многомерное пространство данных. Эта система строит аппроксимирующие двумерное многообразие. Изменение коэффициентов упругости системы позволяет пользователю переключаться от совершенно неструктурированной кластеризации методом динамических ядер к многообразиям близким к линейным многообразиям главных компонент. В промежуточном диапазоне значений коэффициентов упругости, система эффективно аппроксимирует некоторое нелинейное многообразие. Данный подход основывается на аналогии с механикой: главное многообразие, проходящее через «середину» данных, может быть представлено как упругая мембрана или пластинка [8].

2.6. Кластеризация митохондриальных геномов методом динамических ядер

Огромное количество информации требует новых путей обработки и использования. Две наиболее важные проблемы заключаются в классификации и кластеризации геномов. Иначе говоря, автоматического определения группы, к которой принадлежит ранее неизвестный геном и группировка последовательностей геномов в виде древовидной структуры в соответствии с их сходством [18].

Всякий частотный словарь отображает геном в 64-мерное метрическое пространство; близость двух геномов задается естественным образом — например, как близость двух точек в той или иной метрике. В данной работе

использовалась Евклидова метрика. Один из 64 триплетов исключался, поскольку сумма всех частот в словаре равна 1. Формально исключить можно любой триплет, однако в данной работе исключался тот, для которого дисперсия, наблюдаемая по анализируемой выборке геномов, является минимальной [16]. Был исключён триплет GCG, так как в исследуемой выборке он имеет наименьшее стандартное отклонение по сравнению с другими триплетами [19]. Для кластеризации методом упругих карт брался только один тип животных, *Chordata*, который содержал в себе пять классов в данной базе: *Actinopterygii*, *Amphibia*, *Aves*, *Mammalia*, *Reptilia*. Была построена детальная карта и далее таксономические классы выделили разными цветами. Наблюдалось разделение классов на кластеры (Рисунок 11).

2.7. Анализ кластеризации

Главная задача исследования заключалась в проверке того, насколько близки таксономически геномы, близкие и по структуре. Для этого производилась кластеризация с постепенным увеличением классов кластеризации.

Основной вопрос исследования: существует ли наследование при переходе от классификации на K классов к классификации на $K+1$ класс, т.е. верно ли, что такие переходы осуществляются большими группами. Для целей визуализации полученных результатов наиболее естественным объектом являются графы.

При построении кластеризаций вершинами графа являются выделяемые классы. Рёбрами графа являются связи, показывающие переходы групп организмов из классификации на K классов в классификацию на $K+1$ класс.

В понятии графа существенно, что линии соединяют точки. При это не важно, прямые они или являются криволинейными кривыми дугами, длинные они или короткие.

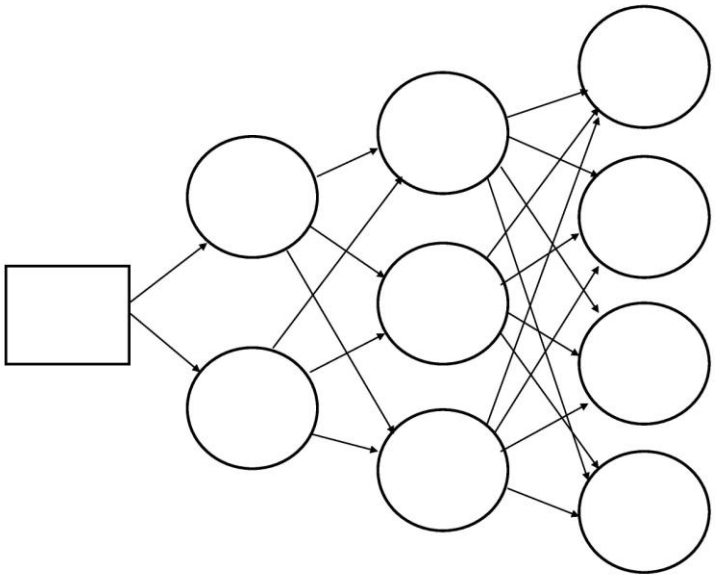
Это приводит к определению графа как абстрактного математического понятия. Рассматривается множество V, состоящее из соединённых некоторым образом точек. Пусть V — это множество вершин и элементы $v \in V$ — вершинами. Граф

$$G = G(V) \qquad (6)$$

с множеством вершин V есть некоторое семейство сочетаний или пар вида

$$E = (a, b), \quad a, b \in V, \qquad (7)$$

указывающее, какие вершины считаются соединёнными. В соответствии с геометрическими представлениями графа каждая конкретная пара (7) называется ребром графа; вершины a и b называются концами ребра E.



Квадратом обозначена исходная база данных. Круги и квадрат — вершины, стрелки — рёбра графа.
Рисунок 2 – Слоистый полносвязный граф

Поскольку в данной работе строили серию классификаций с увеличением числа классов на 1, постольку нас интересуют переходы из классификации на K классов в классификацию на $K + 1$ класс. Для этого используются такие объекты, как слоистые графы.

Слоистый граф — это такой граф, в котором рёбрами соединены вершины, принадлежащие соседним слоям, и только они. В нашем случае мощность слоя (число вершин в слое) монотонно возрастает с ростом

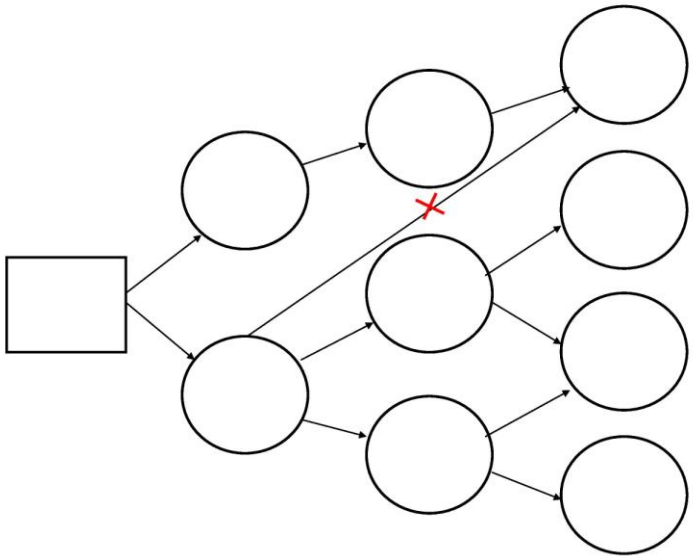
порядка слоя. Вершины слоистого графа разбиты на множества (слои), т.е. слой — это подмножество $G^{(l)}(V)$ вершин графа $G(V), 1 \leq l \leq l^*$ таких, что эти подмножества считаются упорядоченными:

$$G^1(V) \prec G^2(V) \prec \cdots \prec G^{(l-1)}(V) \prec G^l(V) \qquad (8)$$

где символ $\prec$ обозначает отношение упорядочения. Данное отношение не является естественным, но накладывается на множества вершин графа $G(V)$ по соглашению [20 - 21].

Слоистый связный граф — это граф, в котором между двумя вершинами из разных слоёв существует путь (маршрут). Слоистый полносвязный граф — это граф, в котором любая вершина слоя K соединена со всеми вершинами слоя $K + 1$ (Рисунок 2).

При построении кластеризаций результат заранее не был известен. Существовало три возможных варианта графов. Первый вариант — это слоистый полносвязный граф (рисунок 2). Второй вариант, полярный первому — это граф типа дерева. Он представляет из себя связный ацикличный граф. Получение первого варианта говорило бы об отсутствии связи между структурой (частотами триплетов геномов митохондрий) и таксономией их носителя. Под порядком понималось заметное отличие полученного графа от слоистого полносвязного графа. В графе встречаются циклы, в котором какие-то две вершины могут быть соединены разными путями (маршрутами). Если путь между вершинами один, то это говорит об отсутствии циклов (ацикличность). В полученном графе отсутствовала связь между двумя вершинами, находящимися в K слое и $K + 2$ (рисунок 3).



Перечёркнутый указатель перехода говорит об отсутствии связи между K слоем и $K + 2$.

Рисунок 3 – Слоистый граф.

Формально кластеризацию методом динамических ядер можно проводить двумя способами; назовём их условно «сверху вниз» и «снизу вверх». Первый способ состоит в выделении на первом шаге двух кластеров; на следующем — каждый из полученных кластеров также делим на два кластера и так далее, до максимума. При таком делении подразумевается, что при построении графа получится граф типа дерево. Второй способ заключается в последовательном делении исходного множества на $2, 3, \ldots, L$ классов, а затем прослеживается судьба геномов из j-го класса ($1 \leq j \leq R$) при переходе от разбиения на R классов к разбиению на $R - 1$ класс; здесь $L = max\{R\}$. Мы строим исключительно классификацию «снизу вверх».

3. Результаты

3.1. Индексирование базы

Изучалось деление множества геномов методом динамических ядер на два кластера. Каждый геном представлен точкой в 63-мерном пространстве частот триплетов. Выделялись устойчивые группы геномов, относившиеся при делении всегда к одному классу в 500 реализациях из 500.

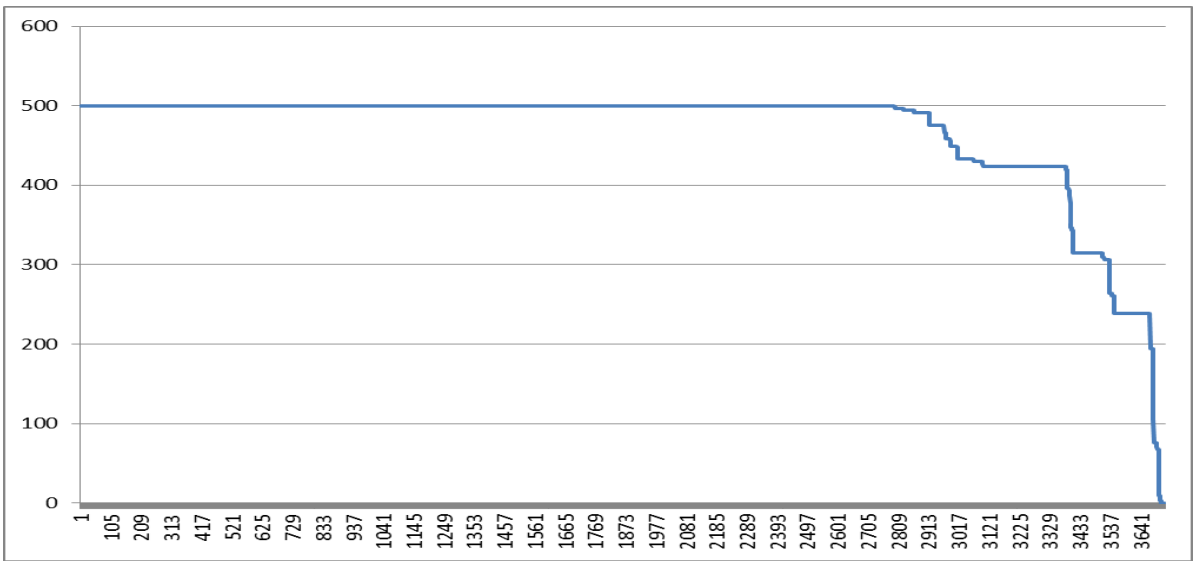


Рисунок 4 – Количество видов устойчиво делящихся при итерациях в исходной базе

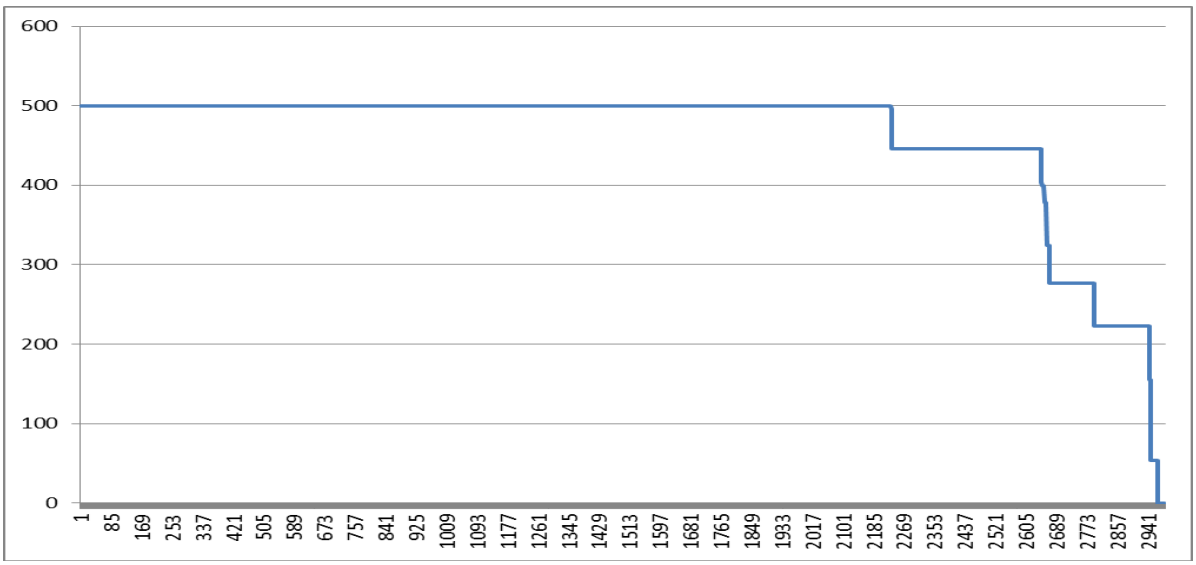


Рисунок 5 – Количество видов устойчиво делящихся при итерациях в индексированной базе

При кластеризации симметрическая разность различалась для каждой абсолютно устойчивой группы. При этом в исходной базе абсолютная группа составляет 71 %, а в индексированной — уже 74 % от всего количества геномов. Мощность множества геномов, составляющая симметрическую разность очищенной базы уменьшилась по сравнению с первоначальной. Число геномов, которые переходят из класса в класс при кластеризации, уменьшилось. Статистическая устойчивость определялась для баз при построении классификации методом динамических ядер. Идея статистической устойчивости состоит в том, что при увеличении величины выборки частота случайного события или среднее значение физической величины стремится к некоторому фиксированному числу. На рисунке 4 представлено разделение видов по количеству реализаций исходной базы данных. На рисунке 5 также представлено разделение видов по количеству реализаций, но уже очищенной базы [22].

3.2. Кластеризация и устойчивость кластеризации

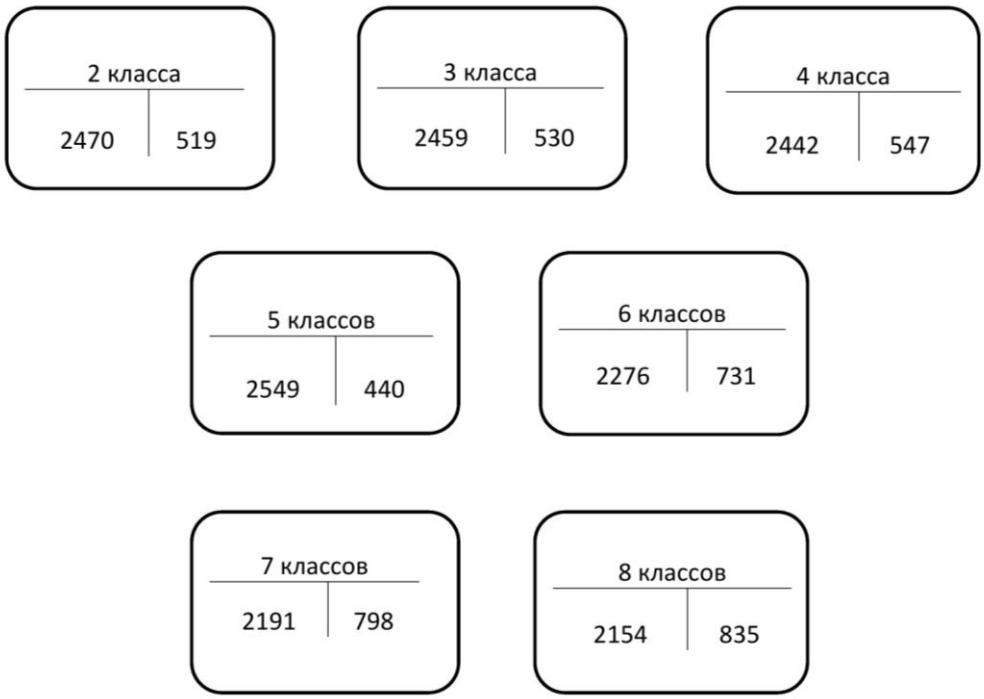


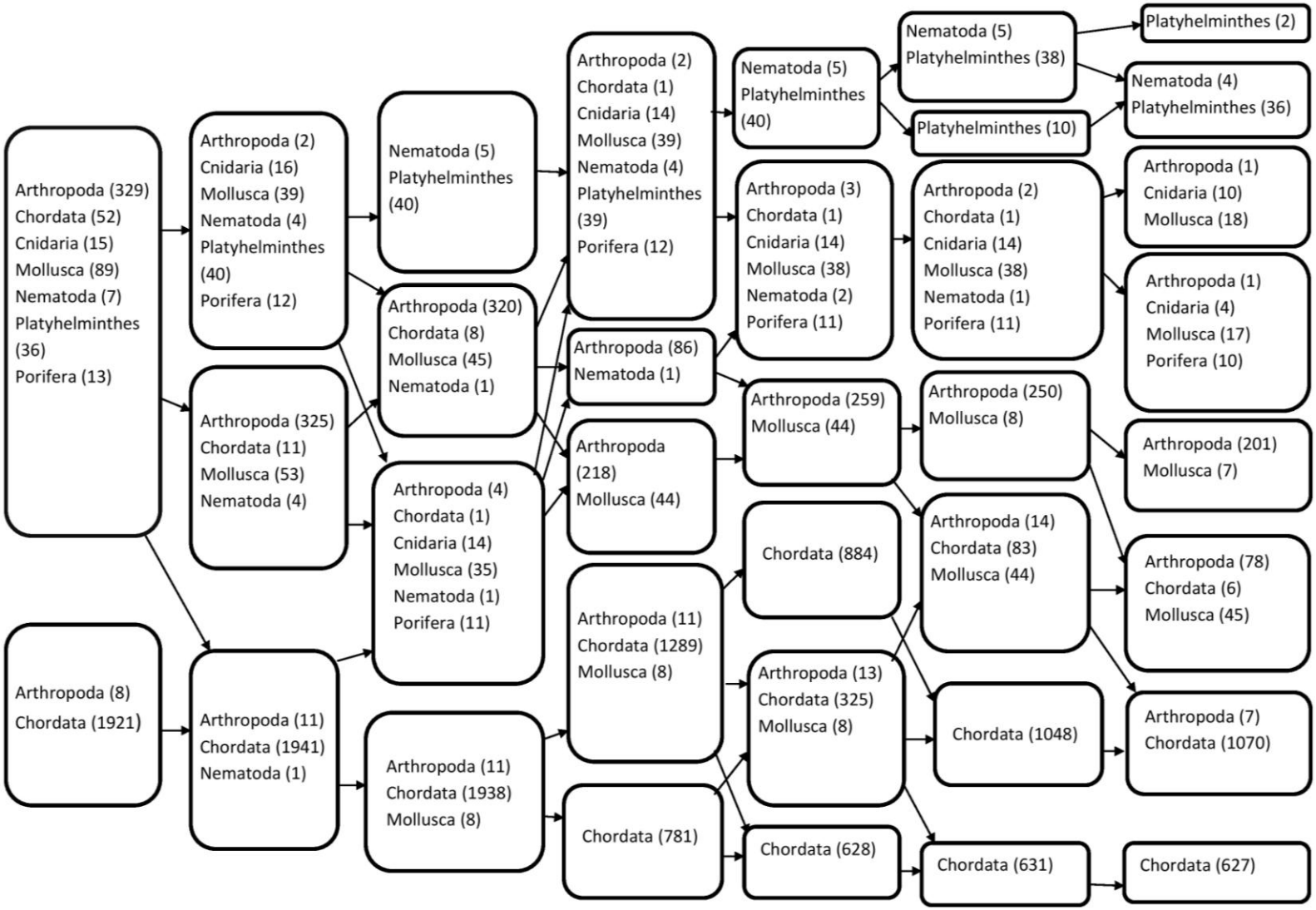
Рисунок 6 – Распределение устойчивых и волатильных видов в базе

Поскольку кластеризация методом динамических ядер зависит от случайного распределения геномов по классам, постольку результирующая кластеризация может оказаться неустойчивой: заметная доля геномов может постоянно менять свою принадлежность к классу. В нашем случае доля волатильных геномов была невелика (число геномов всегда около 500) [15].

При кластеризации поэтапно от 2 до 8 классов было изучено распределение устойчивых и волатильных видов в базе, которое представлено на рисунке 6. Под количеством классов слева указано число устойчиво делящихся видов (то есть геномов, которые при итерациях кластеризации постоянно определялись в один класс), а справа – число волатильно делящихся. При кластеризации волатильные геномы при каждой итерации оказывались в другом классе.

3.3. Распределение таксонов в слоистом графе

Был построен слоистый граф кластеризации по таксономическим типам (рисунок 7). Более подробное распределение видов указано на рисунке 8, который включает в себя таксономические классы.



Число, указанное в скобках — количество геномов в каждом таксоне.

Рисунок 7 – Слоистый граф кластеризации по таксономическим типам

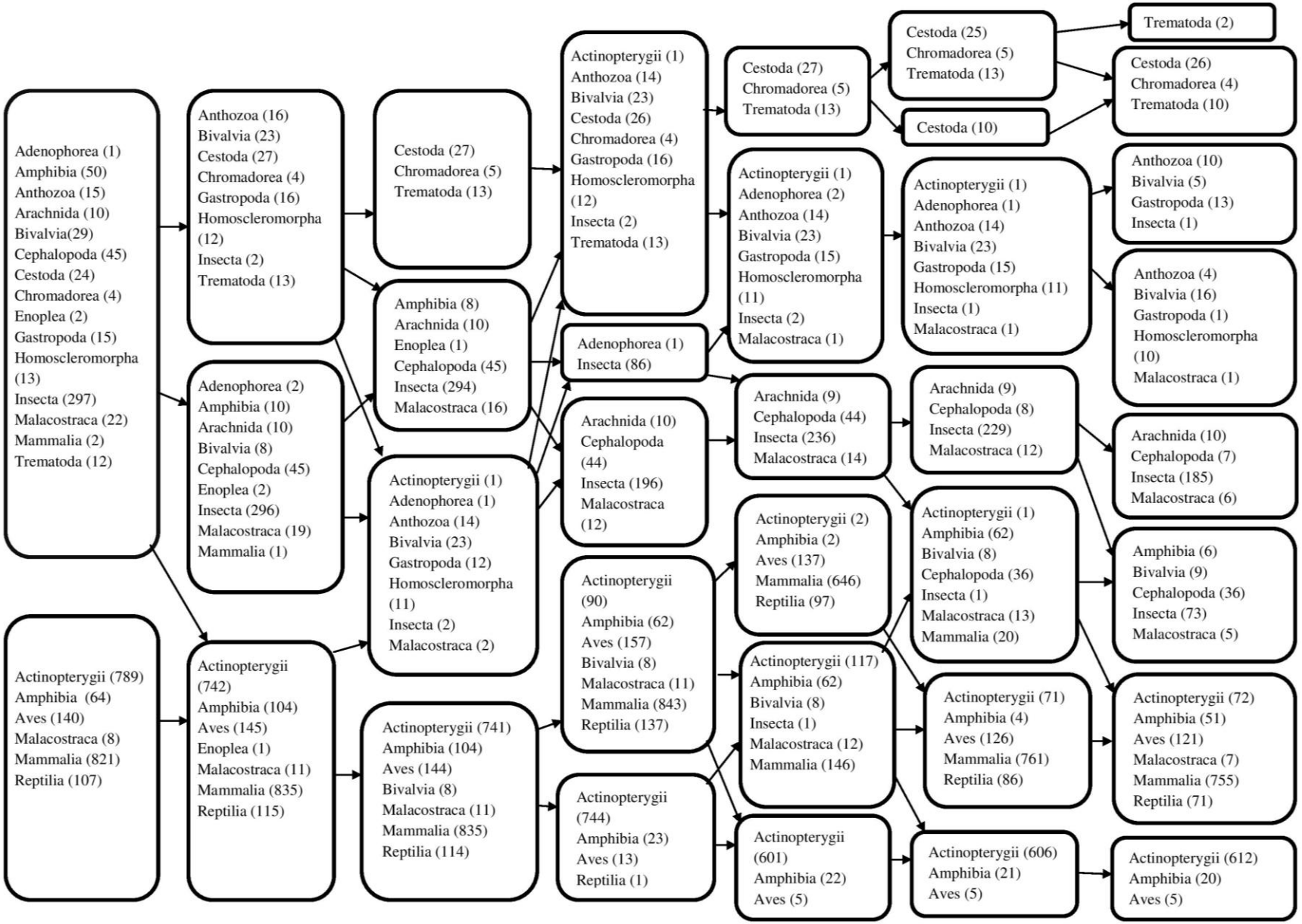


Рисунок 8 – Слоистый граф кластеризации по таксономическим классам

3.4. Упругая карта для типа *Chordata*

Помимо решения прямой задачи, которая заключалась в изучении распределения геномов при кластеризации методом динамических ядер, была решена и обратная задача. Она заключалась в идентификации кластеров через распределение видов методом упругих карт.

На рисунке 9 изображена упругая карта типа *Chordata* 25×25, где виды базы представлены одним цветом.

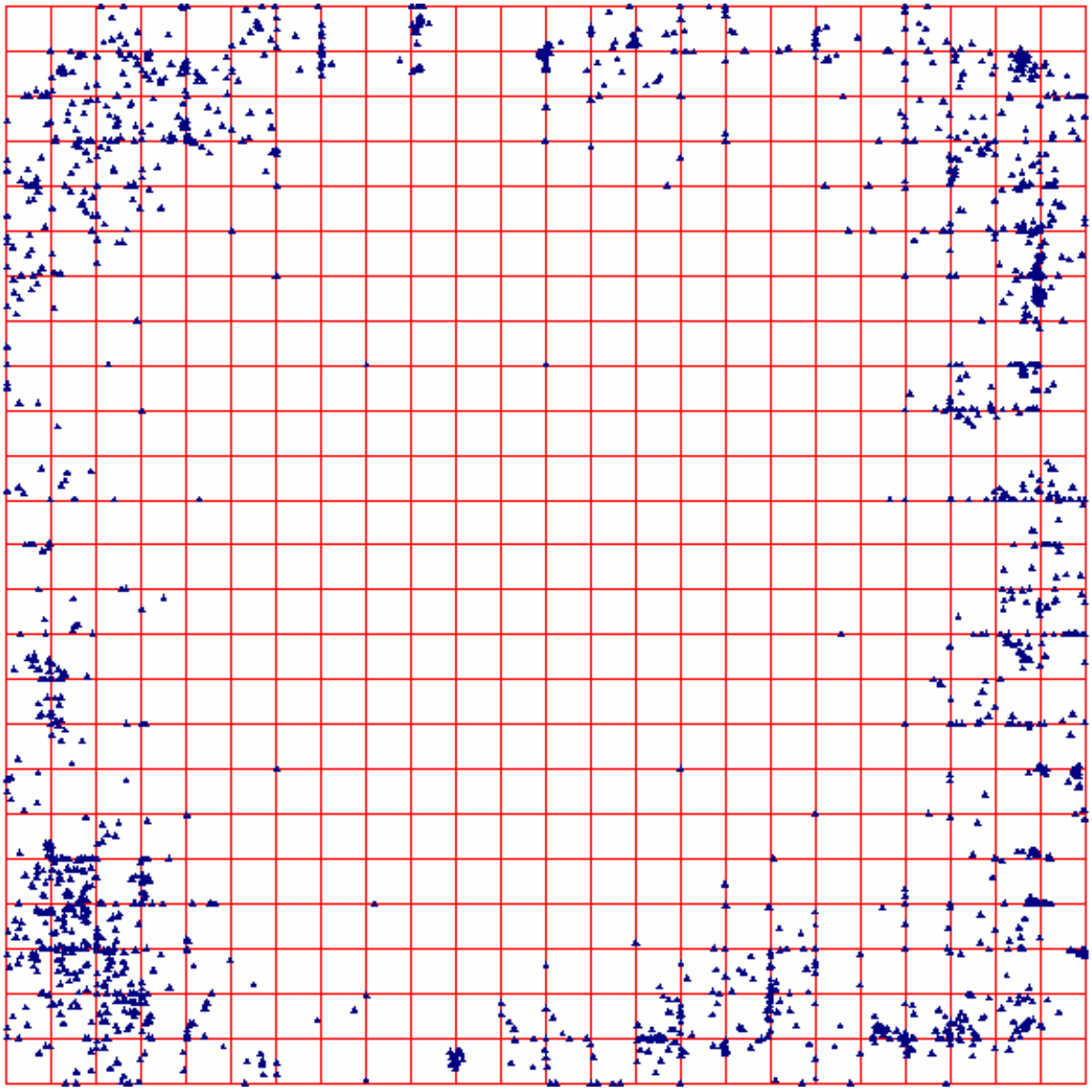


Рисунок 9 – Детальная карта типа *Chordata* 25×25

На рисунке 9 представлена упругая карта типа *Chordata* 25×25 с локальной плотностью.

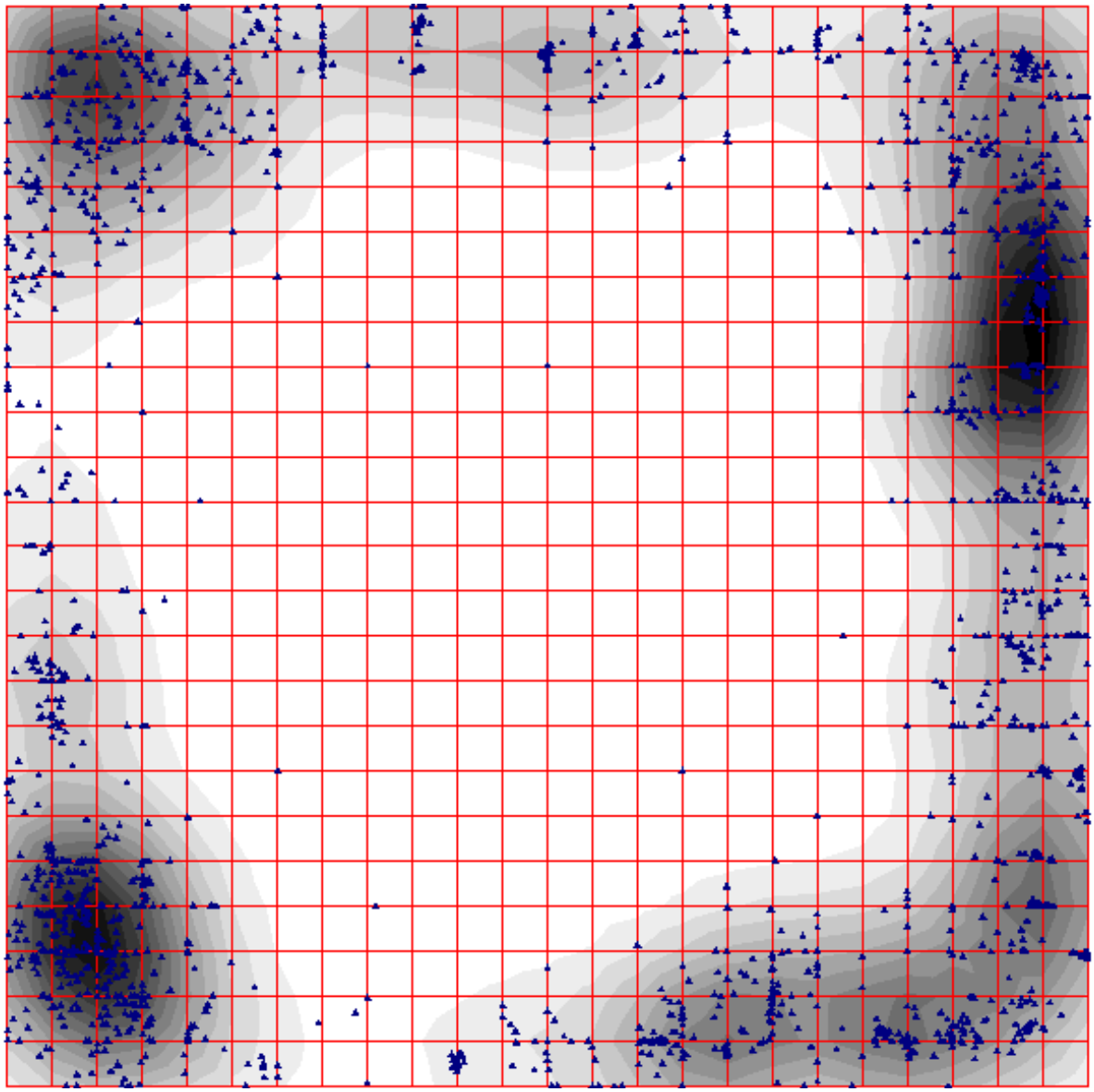


Рисунок 10 – Детальная карта типа *Chordata* 25 × 25 с локальной плотностью

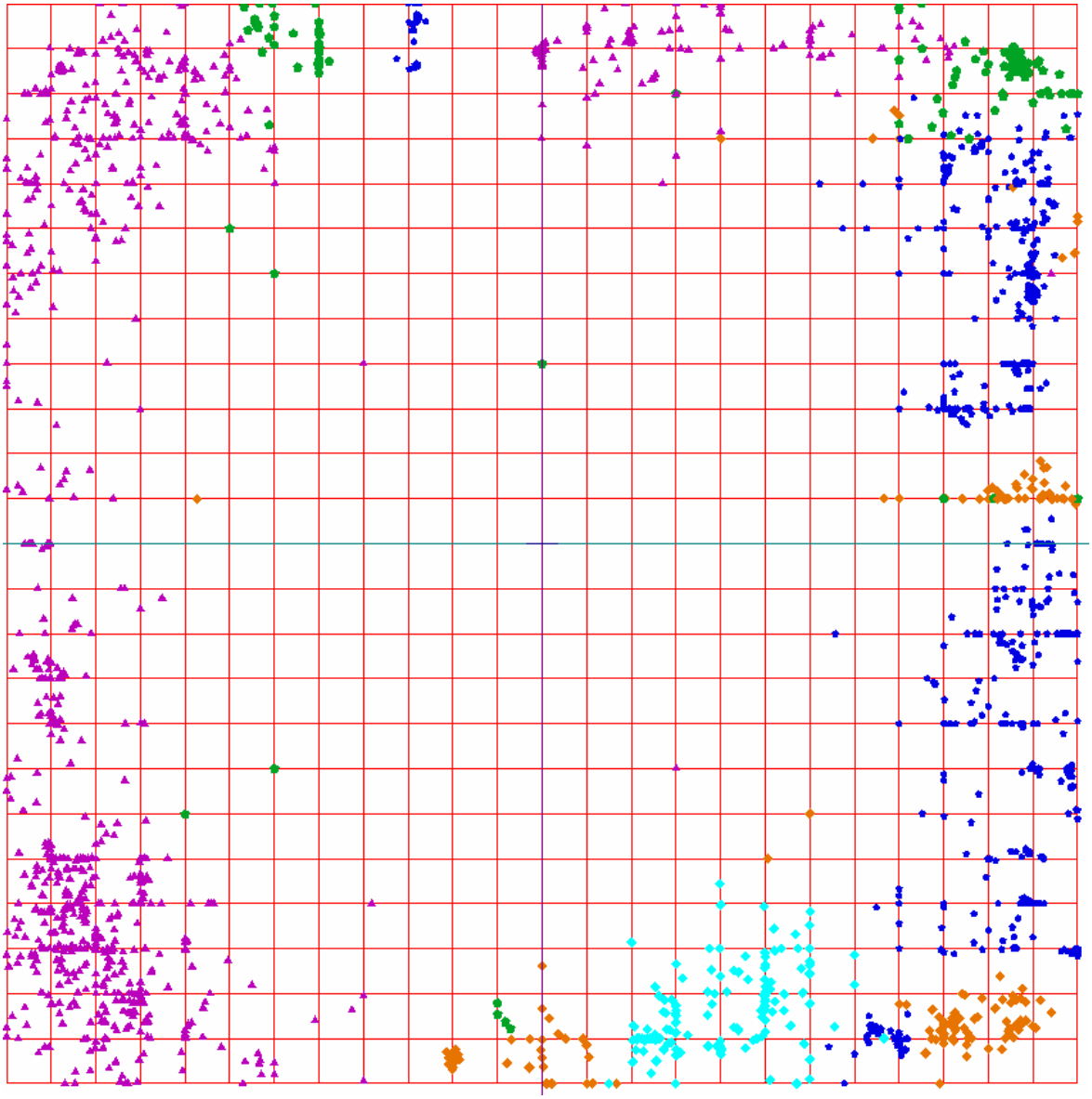


Рисунок 11 – Детальная карта типа Chordata 25 × 25. Классы и соответствующие им цвета:
▲ — Actinopterygii, ♦ — Aves, ⬟ — Amphibia, ⬟ — Mammalia, ♦ — Reptilia.

Рисунок 11 показывает детальную карту распределения геномов для типа *Chordata* с окраской классов. Упругая карта использовалась для визуализации данных. В отдельный кластер отделился только класс *Aves*, остальные образовали от двух (*Actinopterygii*, *Amphibia*) до 4 кластеров (*Mammalia*, *Reptilia*). На карте также видно, что были виды, которые включались не в свой таксономический класс, а также виды, которые не включались ни в один из кластеров. Чаще всего это были представители класса *Amphibia*.

4. Обсуждение

Анализ результатов показывает (рисунок 7 и рисунок 8), что на этапе деления на два кластера таксономически геномы разделились в большинстве на хордовых и беспозвоночных. Некоторое число геномов распределились не совсем точно, что может быть объяснено особенностью используемой базы. Это явление, а точнее особенности соотношения видов с таксономически нетипичными им группами, может быть изучено в последующих работах. К беспозвоночным определились геномы класса *Mammalia* (2 вида) и *Amphibia* (50 видов). В базе существует всего 4 представителя семейства *Erinaceidae* (ежовые). Из них 2 вида — это гимнуры, а другие два — обыкновенный ёж и ушастый ёж. Последние два также относятся к подсемейству *Erinaceinae* (настоящие ежи). Именно эти два вида определились к первому кластеру. Среди класса *Amphibia* распределение можно посмотреть в таблице 6 (в скобках указаны номера классов).

Таблица 6 – Распределение видов таксономического класса *Amphibia* между двумя классами кластеризации

Семейство\Кластеры	Беспозвоночные (1)	Позвоночные (2)
Alytidae	—	2
Ambystomatidae	4	3
Bombinatoridae	—	15
Bufonidae	—	1
Caeciliidae	2	3
Dicroglossidae	—	3
Hylidae	—	1
Hynobiidae	40	4
Ichthyophiidae	—	1
Mantellidae	—	1
Microhylidae	—	1

Pipidae	—	4
Plethodontidae	3	5
Ranidae	—	8
Rhacophoridae	—	2
Rhinatrematidae	—	1
Rhyacotritonidae	1	—
Salamandridae	—	9

Некоторые семейства разделились между двумя кластерами. Преобладающее количество геномов у беспозвоночных имеет семейство *Hynobiidae* (углозубы). К кластеру хордовых определилось 8 видов беспозвоночных, это представители таксономического класса *Malacostraca*.

Таблица 7 – Виды таксономичекого класса Amphibia, попавшие к беспозвоночным

Род	Вид	Класс	Отряд	Семейство
Desmognathus	*fuscus*	*Amphibia*	*Caudata*	*Plethodontidae*
Dicamptodon	*aterrimus*	*Amphibia*	*Caudata*	*Ambystomatidae*
Geotrypetes	*seraphini*	*Amphibia*	*Gymnophiona*	*Caeciliidae*
Hydromantes	*brunus*	*Amphibia*	*Caudata*	*Plethodontidae*
Hynobius	*arisanensis*	*Amphibia*	*Caudata*	*Hynobiidae*
Hynobius	*formosanus*	*Amphibia*	*Caudata*	*Hynobiidae*
Onychodactylus	*fischeri*	*Amphibia*	*Caudata*	*Hynobiidae*
Pachyhynobius	*shangchengensis*	*Amphibia*	*Caudata*	*Hynobiidae*
Ranodon	*sibiricus*	*Amphibia*	*Caudata*	*Hynobiidae*
Rhyacotriton	*variegatus*	*Amphibia*	*Caudata*	*Rhyacotritonidae*

На следующем этапе кластеризации было 3 кластера, распределение произошло в большей степени на 2 кластера беспозвоночных и 1 кластер хордовых. К одному кластеру беспозвоночных определились 1 вид таксономического класса *Mammalia* и 10 видов *Amphibia*.

Эти геномы класса *Amphibia* пришли из класса беспозвоночных на прошлом этапе кластеризации, виды представлены в таблице 7. Вид класса

Mammalia в этом классе кластеризации относится к семейству *Erinaceidae* (ежи).

На третьем этапе кластеризации образовалось 4 кластера, в один из них входят представители только 3-х таксономических классов: *Chromadorea* (нематоды), *Trematoda* (плоские черви) и *Cestoda* (ленточные черви). Если идти по рисунку 8 сверху вниз, то второй клсастер включает в себя 8 представителей таксономического класса *Amphibia*, которые пришли из таблицы 7. Остальные представители этого кластера относятся к беспозвоночным, но в основном довольно высокого порядка (насекомые). В третьем классе появился 1 вид таксономического класса *Actinopterygii* (костные рыбы), остальные виды относятся к беспозвоночным более низкого порядка (моллюски). В четвёртый кластер входят виды хордовых, 8 представителей таксономического класса *Bivalvia*, которые пришли из волатильных видов, 11 видов *Malacostraca* — определились из группы хордовых животных. Совместное движение большей части этого кластера наблюдается уже на трёх этапах деления.

На следующем этапе примечательным оказалось разделение таксономического класса *Insecta* между двумя кластерами. К кластеру с большим содержанием насекомых также принадлежат представители головоногих и высших раков. В первый кластер определились остальные беспозвоночные, 1 вид рыб пришёл из прошлого этапа, где также был один в своём кластере. Этот вид *Tetrabrachium ocellatum* относится к семейству *Antennariidae*. В последних двух кластерах распределились хордовые: последний кластер в основном представляют виды рыб, а верхний кластер представляют виды таксономических классов *Amphibia*, *Aves*, *Mammalia* и *Reptilia*. Также с этой группой перешли представители *Malacostraca* и *Bivalvia*.

Пятый этап деления снова определил представителей 3-х классов червей и большинство представителей рыб в отдельные кластеры. Появился кластер примерно с одинаковым содержанием видов представителей

Amphibia и *Mammalia*. С ними ушли виды *Malacostraca* и *Bivalvia*. Это наталкивает на мысль о том, что виды, притянувшие этих нетипичных для этой группы представителей, находятся именно в этом классе.

Таблица 8 – Виды рыб и птиц в последнем классе шестого этапа деления

Род	Вид	Класс	Отряд	Семейство
Amolops	*tormotus*	*Amphibia*	*Anura*	*Ranidae*
Bombina	*variegata*	*Amphibia*	*Anura*	*Bombinatoridae*
Bombina	*variegata*	*Amphibia*	*Anura*	*Bombinatoridae*
Bombina	*variegata*	*Amphibia*	*Anura*	*Bombinatoridae*
Bombina	*variegata*	*Amphibia*	*Anura*	*Bombinatoridae*
Bombina	*bombina*	*Amphibia*	*Anura*	*Bombinatoridae*
Bufo	*gargarizans*	*Amphibia*	*Anura*	*Bufonidae*
Euphlyctis	*hexadactylus*	*Amphibia*	*Anura*	*Dicroglossidae*
Fejervarya	*cancrivora*	*Amphibia*	*Anura*	*Dicroglossidae*
Hoplobatrachus	*tigerinus*	*Amphibia*	*Anura*	*Dicroglossidae*
Rana	*chensinensis*	*Amphibia*	*Anura*	*Ranidae*
Rana	*chosenica*	*Amphibia*	*Anura*	*Ranidae*
Rana	*dybowskii*	*Amphibia*	*Anura*	*Ranidae*
Rana	*catesbeiana*	*Amphibia*	*Anura*	*Ranidae*
Trichuris	*carnifex*	*Amphibia*	*Caudata*	*Salamandridae*
Trichuris	*cristatus*	*Amphibia*	*Caudata*	*Salamandridae*
Trichuris	*dobrogicus*	*Amphibia*	*Caudata*	*Salamandridae*
Triturus	*karelinii*	*Amphibia*	*Caudata*	*Salamandridae*
Triturus	*marmoratus*	*Amphibia*	*Caudata*	*Salamandridae*
Triturus	*karelinii*	*Amphibia*	*Caudata*	*Salamandridae*
Xenopus	*Silurana*	*Amphibia*	*Anura*	*Pipidae*
Anas	*poecilorhyncha*	*Aves*	*Anseriformes*	*Anatidae*
Anas	*crecca*	*Aves*	*Anseriformes*	*Anatidae*
Anas	*formosa*	*Aves*	*Anseriformes*	*Anatidae*
Anas	*falcata*	*Aves*	*Anseriformes*	*Anatidae*
Anas	*poecilorhyncha*	*Aves*	*Anseriformes*	*Anatidae*

При 7 классах кластеризации образовались 2 кластера червей, 10 видов *Cestoda* отщепились от остальных, и ещё часть пришла из неустойчивых

видов. *Insecta* более менее определились в один кластер вместе с видами *Arachnida*. Вышестоящий над ними кластер в основном состоит из губок, головоногих и двустворчатых моллюсков. Последний класс состоит из рыб, амфибий и 5 видов птиц. Состав амфибий и птиц можно посмотреть в таблице 8.

В последнем этапе отделилось два вида *Trematoda*, в то время как представители *Cestoda* снова сошлись вместе с ещё двумя таксономическими классами червей. Последний кластер остался почти неизменным. Кластер с двустворчатыми и головоногими разделился на две примерно равные группы. *Insecta* образовали два кластера, в одном они объединились с видами *Arachnida*, *Malacostraca* и *Cephalopoda*, а в другом есть ещё представители *Amphibia* и *Bivalvia*.

Существует группа неустойчивых геномов, которая меняется при каждом этапе кластеризации, от слоя к слою. Но и они, в свою очередь, переходили небольшими группами. Колебания в численности видов на рисунках 7 и 8 (указаны в скобочках) объясняется приходом и уходом волатильных геномов в базу устойчивых видов.

Метод упругих карт помог решить обратную задачу, также мы видим, что распределение классов на карте довольно точно соответствует распределению локальной плотности. Более выраженную локальную плотность образовывали кластеры с большим числом видов.

5. Выводы

Показана синхрония в эволюции двух физически независимых изолированных систем (нуклеарный и митохондриальный геномы). При работе с полными митохондриальными геномами она проявляется в неслучайном распределении таксономических категорий внутри каждого слоя, а также в неслучайном перемещении их из слоя в слой.

Показана устойчивость деления. При делении методом *k*-средних большая часть геномов базы устойчива (больше 2000). Кластеризация на три класса и далее показала, что некоторые группы геномов переходят вместе из слоя в слой, что говорит нам об их особенной устойчивости. Также из слоя в слой наблюдается таксономическая закономерность: геномы в большем количестве принадлежат к одной или же смежным таксономическим группам.

Список использованных источников

1. Гистология: Учебник / Ю. И. Афанасьев, Н. А. Юрина, Е. Ф. Котовский и др., Под ред. Ю. И. Афанасьева, Н. А. Юриной — 5-е изд., перераб. и доп. — М.: Медицина, 2002. — С. 64-67.
2. Б. Льюин / Гены; пер. 9-го англ. изд. — М. : БИНОМ. Лаборатория знаний, 2012. — С.10.
3. И. А. Захаров-Гезехус (2014) Цитоплазматическая наследственность // Вавиловский журнал генетики и селекции, Т. 18, № 1. УДК 575.133
4. Gaston H Gonnet. Surprising results on phylogenetic tree building methods based on molecular sequences // BMC Bioinformatics, 2012, 13:148.
5. S. E. Mazzeo, K. S. Mitchell, C. M. Bulik, T. Reichborn-Kjennerud, K. S. Kendler, M. C. Neale. Assessing the heritability of anorexia nervosa symptoms using a marginal maximal likelihood approach // Psychological Medicine / Volume 39 / Issue 03 / 2009, pp 463-473.
6. Alex Zelter, Mojca Bencina, Barry J. Bowman, Oded Yarden, Nick D. Read. A comparative genomic analysis of the calcium signaling machinery in *Neurospora crassa*, *Magnaporthe grisea*, and *Saccharomyces cerevisiae* // Fungal Genetics and Biology /Volume 41 / 2004, pp 827-841.
7. Esbensen K. Multivariate Data Analysis — in Practice // CAMO Process AS / 5-th Edition, 2002.
8. Зиновьев А. Ю. Визуализация многомерных данных. / Красноярск: Изд-во КГТУ, 2000. — С. 168.
9. Мандель И.Д. / Кластерный анализ. — М.: Финансы и статистика. — 1988. — С. 36-37.
10. Jure Leskovec, Anand Rajaraman, Jefrey D. Ullman. Mining of Massive Datasets // Cambridge University Press / 2014, pp 241-250.
11. Горбань А. Н, Попова Т. Г, Садовский М. Г. Классификация нуклеотидных последовательностей по частотным словарям обнаруживает связь между их структурой и таксономическим положением организмов. // Журнал общей биол. 2003. т.64, № 5. С. 16-21.
12. Gorban A.N., Popova T.G., Sadovsky M.G. Classification of symbol sequences over their frequency dictionaries: towards the connection between structure and natural taxonomy // Open Syst. & Information Dyn., 2000, v.7, № 1, pp.1-17.
13. Sadovsky M.G., Shchepanovsky A.S., Putintzeva Yu.A. Genes, Information and Sense: Complexity and Knowledge Retrieval // Theory in Biosciences, 2008, vol. 127, pp. 69-78.

14. Fredrik Lindsten, Henrik Ohlsson, Lennart Ljung. Just Relax and Come Clustering. A Convexification of k-means Clustering // Technical report from Automatic Control at Linköpings universitet (Linköping University), 2011.
15. Горбань А. Н., Россиев Д. А. Нейронные сети на персональном компьютере. / Новосибирск: Наука. 1996. С. — 275.
16. Садовский М. Г., Чернышова А.И. Выявление связи структуры и таксономии геномов хлоропластов методом динамических ядер. // Фундаментальные исследования. № 11-3 / 2014. С. 545-549.
17. ViDaExpert v 1.2 [Электронный ресурс] : Institute of Curie // Andrei Zinovyev. — Режим доступа: http://bioinfo-out.curie.fr/projects/vidaexpert/.
18. A. Tomovic, P. Janicic, V. Keselj *n*-Gram-based classification and unsupervised hierarchical clustering of genome sequences // Computer methods and programs in biomedicine, V.81 (2006), pp. 137-153.
19. Sadovsky M. G. 2014. Evidence for strong co-evolution of mitochondrial and somatic genomes. // arXiv.org
20. О. Оре Теория графов / М.: Наука. Гл. ред. физ.-мат. лит., 1980. — Гл. 1. С. 11-22.
21. Amos Fiat, Dean P. Foster, Howard Karloff, Yuval Rabani, Yiftach Ravid, Sundar Vishwanathan. Competitive Algorithms for Layered Graph Traversal // SIAM Journal on Computing, 1998.
22. В.С. Федотова О популяционной геномике некоторых видов животных / В. С. Федотова, М. Г. Садовский // Сборник материалов Международной конференции студентов, аспирантов и молодых ученых «Проспект Свободный-2015» / 15-25 апреля, 2015. — С. 52-53.