Федеральное государственное автономное образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
Институт фундаментальной биологии и биотехнологии
Кафедра биофизики

УТВЕРЖДАЮ:
заведующий кафедрой

_____ В. А. Кратасюк
“___” _____ 2021 г.

БАКАЛАВРСКАЯ РАБОТА

03.03.02 Физика

СВЯЗЬ СТРУКТУРЫ И ФУНКЦИИ ГЕНОВ ТРАНСПОРТНЫХ РНК ХЛОРОПЛАСТОВ ГОЛОСЕМЕННЫХ РАСТЕНИЙ

Руководитель: _____ д.ф.-м.н., проф. М. Г. Садовский
дата, подпись; уч.степень, должность

Выпускник: _____ Т. О. Шпагина
дата, подпись

Красноярск 2021

РЕФЕРАТ

Выпускная квалификационная работа по теме «Связь структуры и функции генов транспортных РНК хлоропластов голосеменных растений» содержит 51 страницу текстового документа, 4 приложения, 32 использованных источника, 22 рисунка, 7 таблиц.

СТРУКТУРА, ФУНКЦИЯ, НУКЛЕОТИДНЫЕ ПОСЛЕДОВАТЕЛЬНОСТИ, КЛАСТЕРИЗАЦИЯ, УПРУГИЕ КАРТЫ, ГЕНОМЫ ХЛОРОПЛАСТОВ, ТРАНСПОРТНЫЕ РНК, ГОЛОСЕМЕННЫЕ РАСТЕНИЯ

Цель работы — анализ распределения генов тРНК голосеменных растений в пространстве частот триплетов.

Выявление связи между структурой генов (триплетным составом), функции ими определяемой (вид аминокислотного остатка, переносимого в клетки) и таксономией соответствующего гена является классической в молекулярной биологии и биоинформатике. Эта связь анализировалась на примере генов тРНК хлоропластов голосеменных растений, кодирующих 21 аминокислоту, включая формил-метионин. Из полных геномов хлоропластов 145 видов голосеменных извлекались последовательности генов тРНК. Далее база анализировалась по частоте использования кодонов. Нуклеотидные последовательности генов преобразовывалась в частотные словари триплетов с шагом рамки считывания $t = 1$ и $t = 3$. Далее проводилась кластеризация словарей в пространстве частот триплетов нелинейными методами по видам тРНК, синонимичным антикодонам для аминокислот, свойствам тРНК и аминокислот, кодируемых генами.

Ключевой вопрос данной работы звучит так: могут ли тРНК, ответственные за транспортировку специфических аминокислотных остатков, проявлять общие структурные паттерны? Обнаружено, что при кластеризации обоих типов словарей, образуются функционально специфичные кластеры. Результаты данной работы показывают, что для случая генов тРНК не наблюдается связи между составом кластеров и видовым составом организмов. Выявлено преобладание функции над таксономией.

Содержание

1 Введение **5**

2 Обзор литературы **8**

2.1 Геномы хлоропластов голосеменных 8

2.2 Транспортные РНК хлоропластов 10

2.3 Предпочтение кодонов 13

3 Материалы и методы **15**

3.1 Генетический материал и анализ базы 15

3.2 Частотные словари 17

3.3 Метод главных компонент и сопряженные методы кластеризации/классификации 19

3.4 Метод упругих карт 20

4 Результаты **23**

4.1 Краткий обзор исследованной базы генов 23

4.2 Кластеризация словарей методом упругих карт 23

4.3 Кластеризация по синонимичным антикодонам 26

5 Обсуждение **31**

5.1 Краткий обзор полученных результатов 31

5.2 Возможная связь между кластеризацией генов тРНК по частотам триплетов и биохимическими свойствами соответствующих им аминокислот 32

Заключение **35**

Список сокращений **36**

Приложение А **37**

Приложение Б **45**

Приложение В **46**

Приложение Г **47**

Список использованных источников **48**

1 Введение

Исследование связи структуры, нуклеотидных последовательностей, функций, которые в них закодированы, и таксономии носителей этого генетического материала является важной задачей современной молекулярной биологии, биофизики, биоинформатики. Наибольший интерес представляет комплексный подход к изучению связи данных биологических свойств. Настоящая работа посвящена такому анализу на примере генов транспортных РНК хлоропластов. С точки зрения данной работы хлоропласты представляют собой очень удобный объект: все они однородны по своей функции. Это позволяет исключить из анализа различия в функциях используемого геномного материала.

Актуальность настоящей работы обусловлена как задачами анализа большого количества разнообразных геномных данных, результатов секвенирования, для оптимизированного выделения отличительных характеристик организмов, так и новыми возможностями в исследовании связи структуры и функции, открывающимися на больших массивах данных. Геномы хлоропластов могут содержать важную информацию о механизме эволюции голосеменных, поэтому используются в эволюционных и филогенетических исследованиях. Транспортные РНК принимают непосредственное участие в экспрессии генов и могут влиять на аминокислотный профиль организма. Они представляют интерес, поскольку не были изучены статистическими методами с использованием больших массивов данных.

Объектом настоящей работы является связь между структурой и функцией генов транспортных РНК хлоропластов голосеменных.

Целью данной работы является выявление связи между триплетным составом нуклеотидной последовательности генов тРНК хлоропластов голосеменных, их таксономией и функциями этих генов.

Для достижения данной цели были поставлены следующие **задачи**:

1) Создать из полногеномных последовательностей базу генов тРНК и проанализировать её;
2) Построить частотные словари данных последовательностей и провести

кластеризацию словарей различными методами кластеризации и визуализации;

3) Проанализировать распределение словарей по кластерам с точки зрения функционального и таксономического состава.

Работа докладывалась на следующих конференциях:

- 56-я международная научная студенческая конференция (МНСК 2018), Новосибирск, устный доклад;
- X Международная конференция «Dynamical Systems Applied to Biology and Natural Sciences» (DSABNS), Неаполь, стендовый доклад;
- 7th International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO), Гранада, устный доклад;
- 28-й Всероссийский семинар «Нейроинформатика, её приложения и анализ данных», 27 сентября 2019, Красноярск, устный доклад;
- XI международная конференция «Dynamical Systems Applied to Biology and Natural Sciences» (DSABNS), Тренто, стендовый доклад;
- 8th International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO), Гранада, устный доклад;
- Международная конференция студентов, аспирантов и молодых ученых «Проспект Свободный — 2021», Красноярск, устный доклад.

Результаты работы опубликованы в следующих научных журналах и сборниках научных мероприятий:

- Колесникова А.И. Выявление связи тринуклеотидного состава генов и таксономии их носителей на примере генов митохондрий некоторых грибов / Колесникова А.И., Федотовская В.Д., Шпагина Т.О. // Материалы 56-й Международной научной студенческой конференции (МНСК). — 2018. — Vol. 56. — Стр. 18;
- Колесникова А.И. Влияние функциональных различий сильнее влияния таксономических различий для генов семейства atp митохондрий грибов / Колесникова А.И., Федотовская В.Д., Шпагина Т.О., Садовский М.Г. // Моделирование неравновесных систем. / Материалы XXI Всероссийского семинара — 2018. — Vol. 21. — Стр. 49–54;

- Федотовская В. Д. О соотношении влияния функциональных и таксономических различий для генов семейства atp митохондрий некоторых грибов / Федотовская В.Д., Шпагина Т.О., Колесникова А.И., Садовский М.Г. // Нейроинформатика, ее приложения и анализ данных. / Материалы XXVII Всероссийского семинара — 2019. — Vol. 27. — Стр. 110–115;
- Тезисы доклада «Распределение генов митохондрий грибов в пространстве частот триплетов», сборник материалов конференции «Системная биология и биоинформатика» 12-й Международной школы молодых ученых, Ялта, Севастополь, 14-20 сентября 2020 года, индексируется в РИНЦ;
- Sadovsky M., Fedotovskaya V., Kolesnikova A., Shpagina T., Putintseva Y. Function vs. Taxonomy: The Case of Fungi Mitochondria ATP Synthase Genes // Lecture Notes in Computer Science, Springer Verlag. — 2019. — Vol. 11465 — Pp. 335–345;
- Fedotovskaya V. Kolesnikova A., Shpagina T., Putintseva Y., Sadovsky M. Function Overcomes Taxonomy: Case of ATP Genes of Fungi Mitochondria // Tenth International Conference Dynamical Systems Applied to Biology and Natural Sciences: Book of Abstracts. — 2019. — Vol. 10. — Pp. 176–176;
- Fedotovskaya V. Kolesnikova A., Shpagina T., Sadovsky M. The Distribution of Fungal Mitochondrial ATP Genes in Amino Acids Space // 11th International Conference Dynamical Systems Applied to Biology and Natural Sciences: Book of Abstracts. — 2020. — Vol. 11. — Pp. 215-216;
- Fedotovskaya V., Sadovsky M., Kolesnikova A., Shpagina T., Putintseva Y. Function vs. taxonomy: further reading from fungal mitochondrial ATP synthases // Lecture Notes in Computer Science, Springer Verlag. — 2020. — Vol. 12108 — Pp. 438–444.

2 Обзор литературы

2.1 Геномы хлоропластов голосеменных

Голосеменные растения являются жизненно важными компонентами наземных экосистем и имеют важное экономическое и экологическое значение. Согласно последней классификации виды голосеменных делятся на восемь порядков, 12 семейств, 84 рода и более 1000 видов. Голосеменные включают гинкго, саговники, хвойные деревья и гнетофиты, которые выращиваются в лесах как важные породы древесины [4, 25].

Хлоропласты являются пластидами, реализующими фотосинтез и различные метаболические пути. Хлоропласты имеют собственный геном, кодирующий около 100 белков, и наследуются по отцовской линии у большинства голосеменных растений [5, 18, 21]. Как и у прокариот, гены в ДНК хлоропластов организованы в опероны. ДНК хлоропластов имеют кольцевую форму и обычно составляют 120 000–170 000 пар оснований, что составляет около 30–60 микрометров в длину при массе около 80–130 миллионов дальтон [2]. Кроме того, генетический код органелл, в том числе хлоропластов, отличается от генетического кода ядерной ДНК [17].

Согласно теории симбиогенеза, хлоропласты (а также митохондрии) являются эндосимбионтами, которые произошли от сине-зеленых водорослей. Они потеряли свою самостоятельность и приспособились к жизни внутри клетки. В результате клетки-хозяева стали применять к своим нуждам способность эндосимбионтов к фотосинтезу. С течением времени генетический материал из хлоропластов перемещался в ядро, поэтому сейчас хлоропласты уже не могут существовать отдельно от клетки. Доказательствами эндосимбиотической теории могут служить следующие факты. Во-первых, хлоропласты — двумембранные организмы. Они также размножаются бинарным делением, что отличает их от других органелл клетки. Во-вторых, хлоропласты имеют свой аппарат реализации генетической информации. При этом рибосомы имеют константу седиментации 70S, что делает их рибосомами прокариотического типа [28].

Геном — совокупность наследственного материала и состоит из генов и межгенных участков. Этот материал содержится не только в ядре клетки, но и в хлоропластах — особых пластидах фотосинтезирующих эукариот. Геномы хлоропластов могут содержать важную информацию о механизме эволюции голосеменных, поэтому применяются в эволюционных исследованиях.

Весь геном — это последовательность, состоящая из четырех азотистых оснований, расположенных в определенном порядке: A — аденин, C — цитозин, G — гуанин, T — тимин. Ген — единица наследственной информации, отвечающая за формирование у организма какого-либо свойства, выполнения функции на молекулярном уровне; как правило, один ген кодирует какой-нибудь белок. Структуры, отвечающие за передачу наследственной информации, называются экзонами — кодирующими последовательностями. Также в генах существуют ничего не кодирующие участки — интроны.

Наследственная информация, хранящаяся в хлоропластах и ядре не существует полностью независимо. Со временем многие части генома хлоропласта были перенесены в ядерный геном хозяина. Этот процесс называется переносом эндосимбиотического гена. Таким образом прослеживается связь и сходство генетического материала ядра и хлоропластов [17]. Из примерно трех тысяч белков, обнаруженных в хлоропластах, около 95 % кодируются ядерными генами. Многие белковые комплексы хлоропласта состоят из субъединиц как генома хлоропласта, так и ядерного генома хозяина. В результате синтез многих белков должен координироваться хлоропластом и ядром. Хлоропласт в основном находится под ядерным контролем, хотя хлоропласты также могут передавать сигналы, регулирующие экспрессию генов в ядре; этот процесс называется ретроградной передачей сигналов. Таким образом, прослеживается связь и сходство генетического материала ядра и хлоропластов. В хлоропласт гены переносятся намного реже, известно всего несколько таких случаев, например, перенос генов из ДНК хлоропластов в ядерный геном наземных растений. У наземных

растений около 11–14 % ДНК в их ядрах можно проследить до хлоропластов [20].

Несмотря на процесс образования новых видов организмов, эволюция может поддерживать сохранение некоторых генов и соответственно белковых последовательностей. Консервативность гена выражается в сохранении его функции в процессе эволюции в разных видах. Согласно последним исследованиям, для генома хлоропласта характерна консервативность со средней скоростью эволюции $(0{,}2 - 1{,}0) \times 10^{-9}$ на сайт в год, что составляет лишь одну пятую от этого показателя для ядерного генома [6, 7]. Относительная полнота и независимость генома хлоропласта означает, что он может обеспечить ценный материал для исследовательских целей [7].

2.2 Транспортные РНК хлоропластов

Для реализации генетической информации, согласно центральной догме молекулярной биологии, а именно перевода информации, записанной в ДНК, в белковые последовательности, существуют специальные вспомогательные кодирующие молекулы РНК. Различные РНК, такие как матричная, рибосомальная и транспортная, участвуют в процессах транскрипции и трансляции. Транспортные РНК являются важными адапторными молекулами-переносчиками между несущими генетическую информацию нуклеиновыми кислотами и белками в процессе трансляции.

Стоит отметить, что при установлении соответствия кодон-антикодон в процессе декодирования информации взаимодействие между кодоном и антикодоном не абсолютно. Третья позиция в кодоне (первая в антикодоне) подвержена «качанию»: в этом месте не обязательно полностью правильное соответствие кодонов. Это связано с генетическим кодом, точнее с тем фактом, что большинство аминокислот определяются первыми двумя нуклеотидами в триплете. Таким образом, одна и та же тРНК может соединяться с несколькими кодонами [3].

Кратко опишем биохимический процесс работы тРНК. Транспортные РНК связывают активированные по карбоксильной группе аминокислоты

с образованием аминоацил-тРНК и переносят их в рибосому, где осуществляется синтез полипептида. Синтез аминоацилированных тРНК катализируется аминоацил-тРНК-синтетазами при участии АТФ и при участии специфических ферментов для каждой из аминокислот. Каждой белковой аминокислоте соответствует одна или несколько отдельных тРНК, в состав которых входят специфичные только для данной аминокислоты трёхнуклеотидные антикодоны. Данное явление называется вырожденностью генетического кода. В декодирующем центре рибосомы эти участки тРНК распознают комплементарные им кодоны в матричной РНК, связанной с малой субъединицей рибосомы, и вступают с ними в кодон-антикодоновые взаимодействия. При этом их аминоацилированные акцепторные 3′-концы располагаются в пептидилтрансферазном центре рибосомы [16].

Транспортные РНК содержат последовательности менее 100 нуклеотидов, которые складываются во вторичную структуру типа клевер, а затем в L-образную третичную структуру [26]. Вторичная структура тРНК состоит из разных плеч и петель, изображенных на Рисунке 1. Участок, определяющий аминокислоту при трансляции с мРНК в полипептид, называется антикодоном. Именно его нуклеотидный состав определяет аминокислоту в транслированном полипептиде. Каждому уникальному антикодону соответствует одна из аминокислот. Напротив, учитывая вырожденность, одной аминокислоте может соответствовать несколько антикодонов и тРНК. Антикодон тРНК совпадает с аналогичным участком на матрице ДНК с учетом замены тимина на урацил. Вторичная структура и консервативность участков тРНК представлена на рисунке [19]. Как и все нуклеотидные последовательности, антикодон принято прочитывать в направлении от 3′ к 5′ концу. Транспортные РНК хлоропластов голосеменных растений происходят от нескольких общих предков [22]. Особенностью тРНК голосеменных является то, что tRNAIle (изолейцин) кодирует антикодон CAU, который обычно кодируется tRNAMet (метионин) [27].

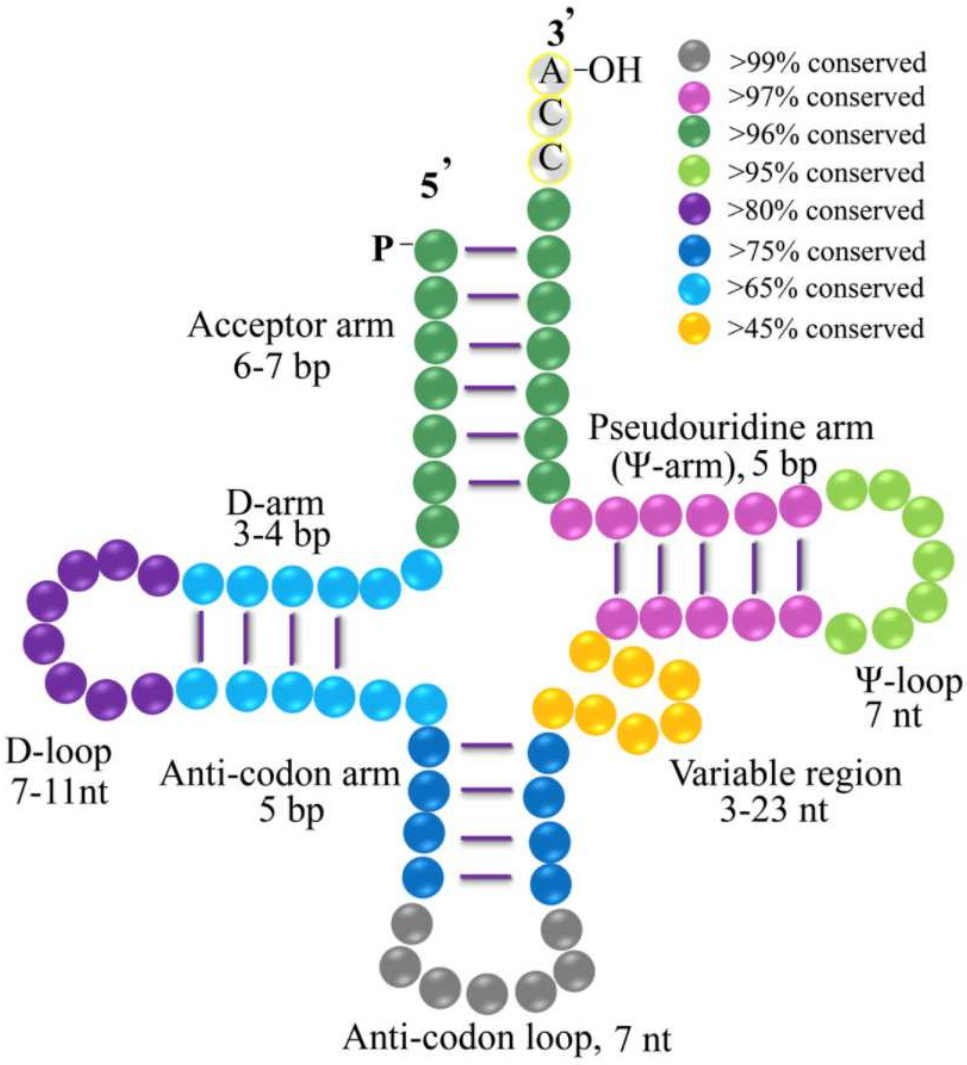


Рисунок 1 — Клеверная листовая структура тРНК голосеменных растений. ТРНК содержит акцепторное плечо (6-7 п.н., темно-зеленый, >96 % консервативных), D-плечо (3-4 п.н., светло-синий, >65 % консервативное), D-петлю (7-11 нуклеотидов, фиолетовый, >80 % консервативно), антикодоновое плечо (5 п.н., темно-синий, >75 % консервативное), антикодоновая петля (7 н., Серый, >99 % консервативных), вариабельная область (3–23 н., Оранжевый, >45 % консервативных), Ψ-плеча (5 п.н., светло-пурпурный, >97 % консервативных) и Ψ-петли (7 нуклеотидов, зеленый, >95 % консервативных). « % Сохранения» означает консервативное соотношение идентичностей оснований в каждой структуре ствола и петли всего набора тРНК голосеменных. Несколько тРНК несут нуклеотиды хвоста C-C-A.

2.3 Предпочтение кодонов

Существует 64 различных кодона (61 кодон, кодирующий аминокислоты, и 3 стоп-кодона), но только 20 различных транслируемых аминокислот (21, включая формилметионин). Как уже было описано ранее, генетический код характеризуется вырожденностью. Кодоны, которые кодируют одну и ту же аминокислоту называют синонимичными. Данное правило относится и к антикодонам. При этом разные кодоны встречаются и используются неравномерно по частоте. Например, в данной работе у голосеменных для кодирования аланина в базе присутствовали 3 тРНК с синонимами антикодонов AGC = 2, GGC = 2, UGC = 98, что говорит о частоте их встречаемости.

В предыдущих исследованиях на модельных организмах было показано, что уровень экспрессии коррелирует со степенью выраженности неравновесия частот кодонов [15]. У разных организмов высокая и низкая частота встречаемости кодонов разная, несмотря на консервативность генетического кода. Гипотеза, согласно которой различные организмы имеют различные предпочтения, называется геномной гипотезой предпочтения кодонов. В разных участках генов, различных геномах и организмах предпочтение кодонов проявляется по-разному под действием эволюционных сил. Данная работа как раз посвящена изучению таких особенностей для генов тРНК голосеменных растений [14].

Хотя механизм выбора смещения частоты кодонов остается спорным, возможные объяснения этого смещения делятся в целом на две теории. Первое объяснение опирается на теорию отбора, в которой смещение кодонов способствует эффективности и/или точности экспрессии белка и, следовательно, подвергается положительному отбору. Модель также объясняет, почему более частые кодоны распознаются более многочисленными молекулами тРНК, а также корреляцию между предпочтительными кодонами, уровнями тРНК и количеством копий генов. Хотя было показано, что скорость включения аминокислот в более частые кодоны значительно выше, чем у редких кодонов, не было показано, что скорость трансляции

напрямую влияет на этот процесс и, следовательно, смещение в сторону более частых кодонов может не быть прямо выгодным. Однако увеличение скорости элонгации трансляции может быть косвенным преимуществом за счет увеличения клеточной концентрации свободных рибосом и, возможно, скорости инициации информационных РНК (мРНК) [15].

Второе объяснение использования кодонов опирается на мутационное смещение — теорию, которая утверждает, что смещение кодонов существует из-за неслучайности мутационных паттернов. Другими словами, некоторые кодоны могут подвергаться большему количеству изменений и, следовательно, приводить к более низким частотам равновесия, также известным как «редкие» кодоны. Дополнительные исследования продемонстрировали, что смещения кодонов можно статистически предсказать у прокариот, используя только межгенные последовательности, что противоречит идее селективных сил на кодирующие области и дополнительно поддерживает модель смещения мутаций. Однако сама по себе эта модель не может полностью объяснить, почему предпочтительные кодоны распознаются более многочисленными тРНК [15].

Неясно, влияет ли использование кодонов на эволюцию тРНК или наоборот. Была разработана по крайней мере одна математическая модель, в которой и использование кодонов, и экспрессия тРНК совместно развиваются в режиме обратной связи (то есть кодоны, уже присутствующие на высоких частотах, повышают экспрессию своих соответствующих тРНК, а тРНК, обычно экспрессируемые на высоких уровнях, повышают частоту их соответствующих кодонов). Однако эта модель, к настоящему времени еще не получила экспериментального подтверждения. Кроме того, эволюция генов тРНК была областью, не привлекавшей большого интереса исследователей.

3 Материалы и методы

3.1 Генетический материал и анализ базы

В качестве генетического материала в данной работе рассматриваются гены транспортных РНК хлоропластов 145 видов голосеменных растений, представленных в открытой базе данных NCBI GenBank в виде файлов нуклеотидных последовательностей. Изначально материал в базе NCBI представлен в виде полногеномных последовательностей, поэтому с помощью программы CLC Genomics Workbench были выделены непосредственно последовательности генов тРНК в количестве 4887. Из-за особенностей аннотирования в выборке содержались повторы генов — один и тот же генетический материал, отсеквенированнный несколько раз разными организациями, и набор последовательностей был отфильтрован до 4531.

Выборка включает в себя 12 семейств, 53 рода и 145 видов голосеменных растений. В Таблице 1 представлено количественное соотношение исследуемой выборки организмов по семействам. Как видно из таблицы, наиболее представлены виды семейств: *Сосновые, Кипарисовые* и *Тисовые*. В Таблице 2 представлено количественное соотношение генов тРНК по аминокислотам. В отличие от ядерного генома, в геномах хлоропластов используются не все кодоны для кодирования определённых аминокислот. В Таблице 3 представлены такие антикодоны из анализа выборки.

Таблица 1 — Состав выборки организмов по семействам. N — число видов.

Семейство	N	Семейство	N
Araucariaceae	4	*Pinaceae*	56
Cupressaceae	32	*Podocarpaceae*	9
Cycas	4	*Sciadopitys*	1
Ephedraceae	4	*Taxaceae*	23
Ginkgoaceae	1	*Welwitschia*	1
Gnetaceae	3	*Zamiaceae*	7

Таблица 2 — Количество генов тРНК в соответствии с аминокислотой, представленное в базе.

тРНК-аминокислота	Количество генов
A	129
C	145
D	148
E	143
F	143
G	258
H	158
I	301
K	141
L	419
M	146
fM	142
N	144
P	275
Q	166
R	371
S	427
T	302
V	271
W	148
Y	148

В исследуемых последовательностях определяющим вид тРНК и аминокислоту, ей соответствующую, участком является антикодон. В файлах последовательностей генов, антикодон также как и вся последовательность, записаны и читаются в направлении от $3'$ к $5'$ концу. Как вид отдельной аминокислоты, переносимой той или иной молекулой тРНК, так и порядок аминокислотного остатка в пептиде определяется взаимным расположением нуклеотидов в гене и иных генетических структурах. Это обстоятельство позволяет рассматривать ген как математический объект — символьную последовательность и применять к ней весь арсенал средств

комбинаторики, статистики и теории вероятностей.

Таблица 3 — Гены тРНК голосеменных с соответствующими аминокислотами и антикодонами, не представленные в базе.

тРНК-аминокислота	Антикодон
A	CGC
G	ACC
I	UAU
K	CUU
L	AAG
L	CAG
L	GAG
N	AUU
P	AGG
Q	CUG
R	CGC
R	UCG
S	ACU
T	AGU

3.2 Частотные словари

Целью данной работы является анализ связей между структурой нуклеотидной последовательности, ее функцией и таксономией. Термин “структура” в контексте данной работы может истолковываться по-разному. Для достижения поставленной цели, необходимо дать определение этому термину. Здесь в качестве структуры нуклеотидной последовательности будем рассматривать частотный словарь триплетов, описанный ниже.

Нуклеотидная последовательность рассматривается как символьная последовательность. Все нуклеотидные последовательности состоят из четырех символов: A, C, G и T. Для любой такой последовательности можно построить частотный словарь подпоследовательностей длиной q. Частотный словарь — это набор слов данной последовательности вместе с инфор-

мацией о их встречаемости.

Для построения частотного словаря необходимо проделать следующее. Рамка считывания длинной q помещается в начало последовательности и перемещается по ней с шагом $t \geqslant 1$. При этом записывается число копий n_ω каждого встретившегося триплета ω. Далее рассчитывается частота каждого q-плета по формуле:

$$f_\omega = \frac{n_\omega}{N}\,, \tag{1}$$

где N — общее число копий триплетов. В настоящей работе использовались два вида шага: 1 и 3.

В данной работе анализировался триплетный состав генов тРНК, т.е. частотный словарь всех триплетов от $\omega_1 = \mathrm{AAA}$ до $\omega_{64} = \mathrm{TTT}$. Таким образом, частотные словари — это точки в 64-мерном пространстве, в котором осями являются частоты всех триплетов. Однако, для наилучшей реализации кластеризации из анализа исключался один триплет вместе с его частотой, поэтому частотные словари рассматривались как точки в 63-мерном пространстве. Это исключение обусловлено тем, что сумма всех частот триплетов равна 1. Указанное пространство является метрическим пространством, в котором можно задать различные функции расстояния. В данной работе использовалась Евклидова метрика, в которой расстояние определяется по теореме Пифагора. В наших обозначениях:

$$d\left(W^{(1)}, W^{(2)}\right) = \sqrt{\sum_{\omega=\mathsf{AAA}}^{\mathsf{TTT}} \left(f_{\omega^{(1)}} - f_{\omega^{(2)}}\right)^2} \tag{2}$$

где $W^{(1)}$ и $W^{(2)}$ — пара частотных словарей, $f_\omega^{(1)}$ и $f_\omega^{(2)}$ — частоты триплета ω пары частотных словарей.

Совокупность всех триплетов и частоту их встречаемости в общем виде обозначим $W_{(q,t)}$. В работе изучались словари $W_{(3,1)}$ и $W_{(3,3)}$. Частотные словари были построены *ad hoc* программой, написанной на языке программирования *Python*.

3.3 Метод главных компонент и сопряженные методы кластеризации/классификации

Кластерный анализ — это группа методов, позволяющих разделить большой объем данных на сравнительно однородные группы и выделить их внутреннюю структурированность. В настоящее время в области биоинформатики порождается огромное количество данных ежегодно. Изучить такой большой объем информации помогают методы кластеризации, которых существует огромное количество. Условно их можно разделить на несколько групп.

В первую очередь, это алгоритмы, основанные на вероятностном подходе. Другими словами, принадлежность той или иной точки к определенному кластеру описывается некоторым вероятностным распределением. Среди этой группы выделяется метод ММЦ (Метод Марковских цепей).

Следующая группа — теоретико-графовые алгоритмы. Другое название этой группы кластеров — иерархическая кластеризация. Суть этих методов заключается в том, чтобы построить иерархию, в которой представлены вложенные кластеры. При этом новые кластеры могут создаваться путем разделения более крупных или, наоборот, объединения более мелких кластеров. Обычно кластеризация, выполненная этими методами, может быть представлена как дендрограмма. Для использования теоретико-графовых алгоритмов необходимо задать меру расстояния. Существует большое количество методов определения расстояния, среди них: методы одиночной, полной, средней связи, центроидный метод, метод Уорда.

Последняя группа методов кластеризации, приведенная в настоящей работе — методы на основе работы искусственного интеллекта. Например, нейронные сети Кохонена.

Не существует универсального алгоритма кластеризации, подходящего во всех случаях. Под разные типы данных и под разные задачи необходимо выбирать разные алгоритмы кластеризации. Это связано с тем, что каждый из алгоритмов имеет свои достоинства и недостатки, которые необходимо учитывать. Кроме того, в зависимости от природы данных и других

факторов может выбираться различное количество кластеров, метрики и другие параметры. Это этого зависит окончательный результат кластеризации.

Анализ данных в биоинформатике осложнен многими факторами. Один из них — тот факт, что как правило, эти данные многомерны. Для анализа таких данных очень часто используют метод главных компонент. Этот метод, являясь линейным, позволяет снизить размерность данных, теряя при этом минимум важной информации. Важной задачей метода главных компонент является аппроксимация данных линейными многообразиями. Для этого вычисляются собственные векторы матрицы ковариаций и ее собственные значения. Этот метод широко применяется в биоинформатике, общественных науках, обработке изображений и т.д. Кроме того, многомерные статистические методы, такие как анализ соответствия и анализ главных компонент, широко используются для анализа вариаций использования кодонов среди генов [1, 9–11, 29, 30].

Предпочтение кодонов (Codon Usage Bias, CUB) — различия в частоте встречаемости синонимичных кодонов в кодирующей ДНК. К основному предназначению предпочтения кодонов относят регуляцию экспрессии генов. Использование кодонов в некодирующих областях ДНК может играть важную роль во вторичной структуре РНК и последующей экспрессии белка, которые могут подвергаться дополнительному селективному давлению. В частности, сильная вторичная структура в сайте связывания рибосомы или кодоне инициации может ингибировать трансляцию, а сворачивание мРНК на $5'$ конце вызывает большое количество вариаций в уровнях белка. Такие методы, как «частота оптимальных кодонов» (Fop), используются для измерения равномерности (равновесности) использования кодонов.

3.4 Метод упругих карт

Метод упругих карт — метод кластеризации и визуализации данных в многомерном пространстве. Служит для сокращения размерности изучаемых данных и является нелинейным методом. Суть метода заключается

в том, что данные в многомерном пространстве проецируются на некую неизогнутую поверхность и отображаются на ней как на карте [9–11]. Такой поверхностью может служить упругая пластина, к которой пружинами прикреплены данные. Система, состоящая из поверхности и прикрепленных к ней данных аппроксимирует многомерные данные двумерным многообразием путем сложного нелинейного преобразования исходно плоского многообразия.

Более подробно упругая карта строится следующим образом. На первом шаге в многомерном пространстве определяется первая главная компонента. Первая главная компонента является направлением вдоль которого в многомерном пространстве наблюдается максимальный разброс данных. Далее строится вторая главная компонента. Она перпендикулярна первой, и направление, вдоль которого она построена, является следующим по величине разброса данных.

На втором шаге на первых двух главных компонентах как на осях строится плоскость. На плоскости необходимо построить проекции всех точек в пространстве. Далее определяется минимальный квадрат, охватывающий все проекции точек. После этого каждая точка соединяется со своей проекцией на плоскости математической пружиной. Эта пружина имеет бесконечную растяжимость, и ее свойства не меняются по мере растяжения.

На третьем шаге жесткий квадрат заменяется гибкой мембраной, способной растягиваться, сжиматься и испытывать деформации сдвига. Далее вся система отпускается. В итоге она достигает такой формы, которая соответствует минимуму общей энергии деформации.

На последнем шаге происходит переопределение точек в результате построения ортогональных проекций для каждой точки. После этого система релаксирует до тех пор, пока снова не примет форму квадрата. Именно этот квадрат и является упругой картой во внутренних координатах [9–11].

При построении упругой карты важно помнить о следующем ограничении. Необходимо сохранить топологию исходного множества. Это озна-

чает, что эластичную мембрану нельзя разрывать и склеивать.

Описанный метод кластеризации подразумевает использование локальной плоскости точек на карте. Для этого каждой точке необходимо поставить куполообразную функцию, например, функцию Гаусса [8]:

$$f_j(r) = A \cdot \exp\left\{-\frac{(r - r_j)^2}{\sigma^2}\right\} . \tag{3}$$

Здесь r_j — координата j–ой точки, A — множитель, одинаковый для всех точек, σ — полуширина этой функции (стандартное отклонение).

После этого необходимо вычислить сумму всех этих функций:

$$F(r) = \sum_{j=1}^{N} \exp\left\{-\frac{(r - r_j)^2}{\sigma^2}\right\} . \tag{4}$$

где N — общее число точек. Для проведения кластеризации и визуализации результатов на упрогой карте использовалось ПО *ViDaExpert*[1] [9–11].

[1] http://bioinfo-out.curie.fr/vidaexpert/

4 Результаты

4.1 Краткий обзор исследованной базы генов

Для использованных в работе 4531 последовательностей генов тРНК голосеменных растений 145 видов были построены частотные словари с шагом 1 и 3. Были вычислены стандартные отклонения в частотных словарях шага 1 и 3 для каждого из 64 триплетов. Далее триплеты с минимальным стандартным отклонением по всей базе последовательностей были исключены из рассмотрения (Таблица 4). Дело в том, что сумма частот всех триплетов равна единице и, следовательно, частоты 64 триплетов не являются нелинейно независимыми. Триплет с наименьшим стандартным отклонением дает наименьший вклад в различимость генов. Именно поэтому он исключался из рассмотрения.

Таблица 4 — Стандартные отклонения

Шаг рамки	σ	Триплет
$t = 1$	0,007331	ACA
$t = 3$	0,004166	ACA

4.2 Кластеризация словарей методом упругих карт

В *ViDaExpert*[2] [9–11]. были построены упругие карты 16×16 для обоих словарей и проведена кластеризация точек в 63-мерном пространстве частот триплетов (Рисунок 2). Полный словарь из 63 триплетов отображает последовательность каждого гена тРНК в 63-мерном пространстве. На рисунке изображены распределения генов тРНК в частотном пространстве по локальной плотности.

Первый важный результат работы заключается в том, что распределение генов тРНК в частотном пространстве очень неоднородно и образует четко идентифицируемые кластеры. Является ли состав кластеров случай-

[2] http://bioinfo-out.curie.fr/vidaexpert/

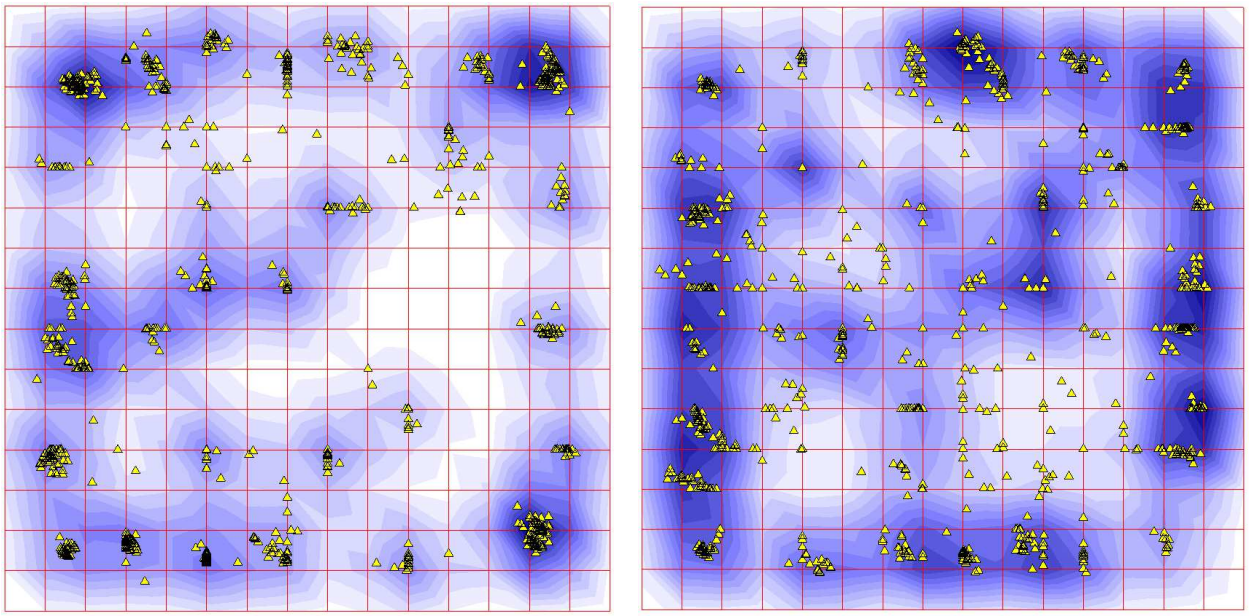


Рисунок 2 — Кластеризация 4531 генов тРНК 145 видов голосеменных растений. Распределение по локальной плотности с контрастным радиусом $r = 0,15$ Слева для шага 1, справа для шага 3. Упругие карты показаны во внутренних координатах .

ным или нет, с точки зрения конкретных генов, входящих в состав кластера и составляет ключевой вопрос работы. Ответ на данный вопрос в целом положителен: кластеры весьма однородны по составу генов тРНК, несущих одну и ту же аминокислоту. Иными словами, гены, кодирующие одну аминокислоту преимущественно концентрируются в одном кластере; этот результат показан на Рисунке 3. Кластеры, содержащие гены, ответственные за конкретную аминокислоту, выделены разным цветом. Количество генов, форма и цвет точек для рисунка указаны в таблице в Приложении Б.

Стоит отметить, что не все гены тРНК аннотированы в базе NCBI с указанием антикодона, поэтому часть из них имеет обозначение только по аминокислоте.

Из анализа частот генов следует, что некоторые гены и антикодоны представлены в исследуемой базе часто, а некоторые — единично. Рассмотрим отдельно распределение таких единичных кодонов. Распределение по

наименее представленным кодонам показано на Рисунке 4. Аннотация к форме, цвету точек указана в приложении.

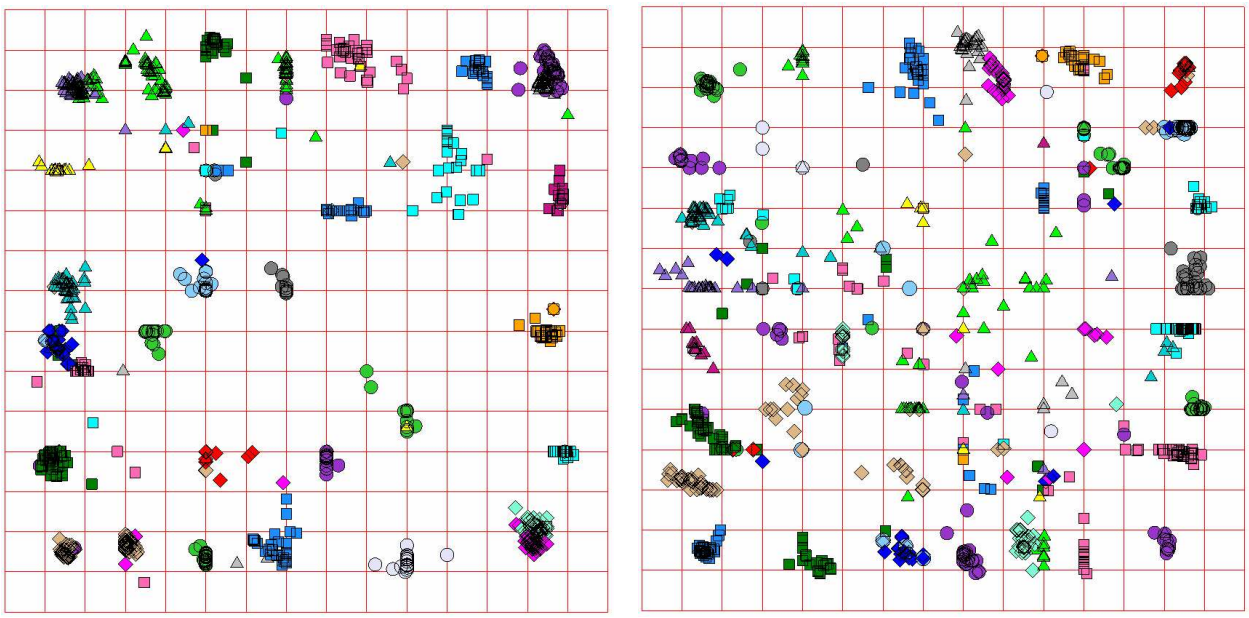


Рисунок 3 — Распределение тРНК по 21 аминокислоте для шага 1 слева, шага 3 справа, различные аминокислоты обозначены разным цветом. Упругие карты показаны во внутренних координатах.

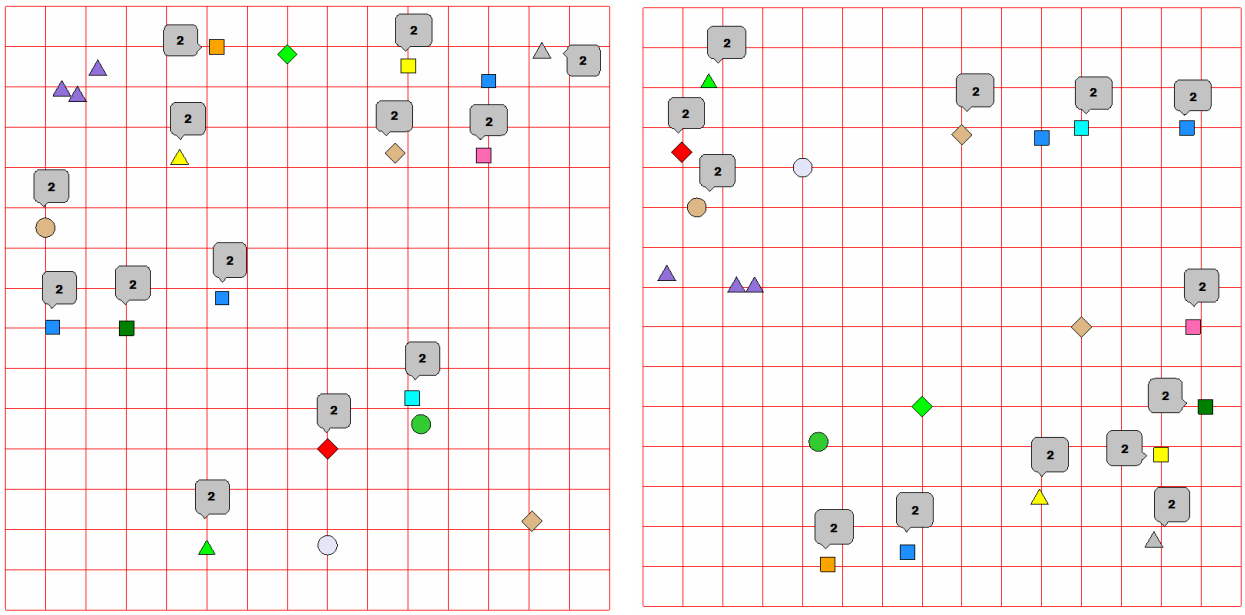


Рисунок 4 — Распределение по наименее представленным кодонам. Шаг 1 слева и шаг 3 справа. Выносками серого цвета указано количество точек.

4.3 Кластеризация по синонимичным антикодонам

За счет вырожденности генетического кода определенные аминокислоты переносятся тРНК с синонимичными антикодонами. В анализируемой базе последовательностей группы тРНК делятся на два вида: представленные одним антикодоном и представленные несколькими синонимичными антикодонами. Синонимичные антикодоны для отдельных аминокислот распределены неравномерно по частоте встречаемости. Далее представлены ответы на вопросы: велико ли число тРНК для единичных антикодонов и как распределены синонимичные по антикодону тРНК в пространстве частот?

Гены с антикодонами с выраженным неравномерным распределением по частоте соответствующих антикодонов в базе не анализировались на расхождение по разным кластерам по принципу синонимичности антикодонов. Такие гены показаны на Рисунке 4. К эти генам относятся аланин (AGC, GGC), цистеин (ACA), аспартат (AUC), глутамат (TTC), глицин (CCC), лизин (AAA), лейцин (AUG), пролин (CGG), аргинин (CCU), серин (AGA, CGA), треонин (CGU), валин (CAC, AAC) и тирозин (AUA).

В то же время гены с антикодонами с соизмеримо равномерным соотношением по частоте встречаемости в базе анализировались на расхождение по разным кластерам по этому признаку. Количественное распределение таких генов указано в таблице в Приложении Г. Это аминокислоты: глицин, изолейцин, лейцин, пролин, аргинин, серин, треонин и валин. На Рисунках 5–12 представлено распределение по синонимичным антикодонам для генов тРНК в частотном пространстве с шагами 1 и 3.

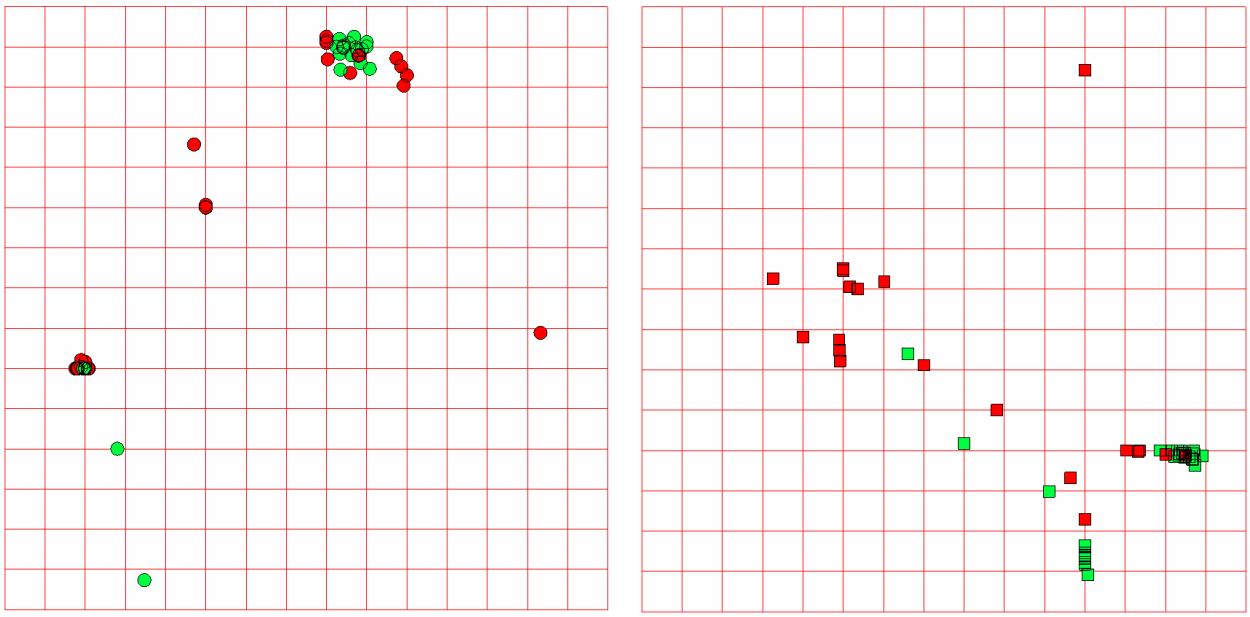


Рисунок 5 — Распределение по синонимичным антикодонам: глицин. Антикодоны: UCC — красным, GCC — зеленым. Шаг 1 слева (круги) и шаг 3 справа (квадраты)

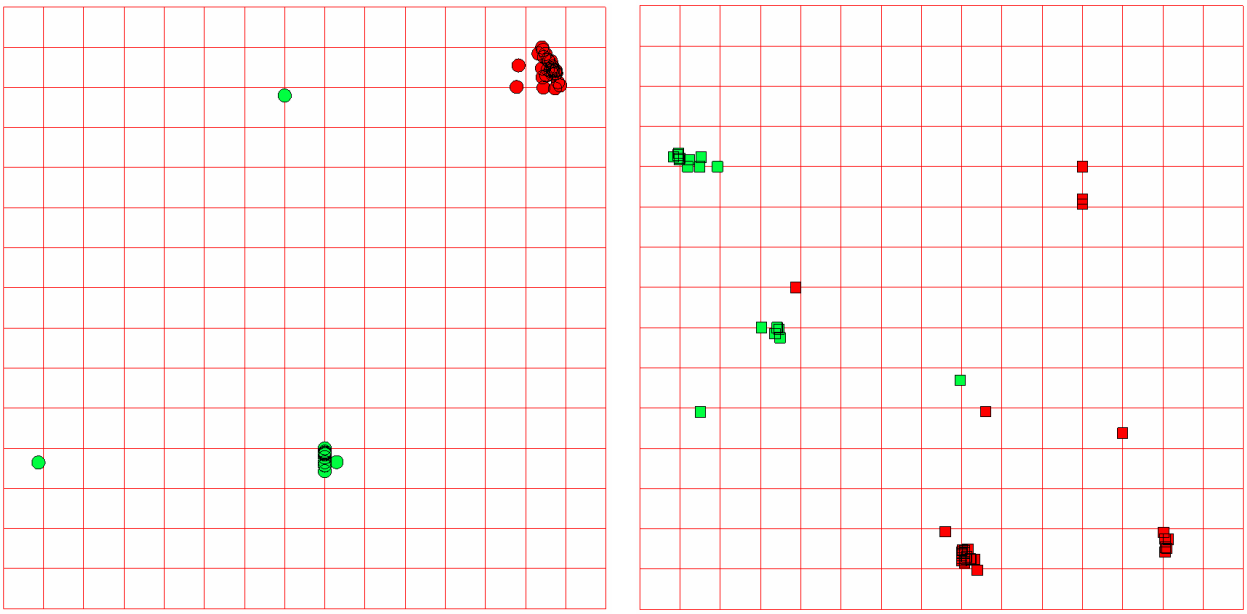


Рисунок 6 — Распределение по синонимичным антикодонам: изолейцин. Антикодоны: CAU — красным, GAU — зеленым. Шаг 1 слева (круги) и шаг 3 справа (квадраты)

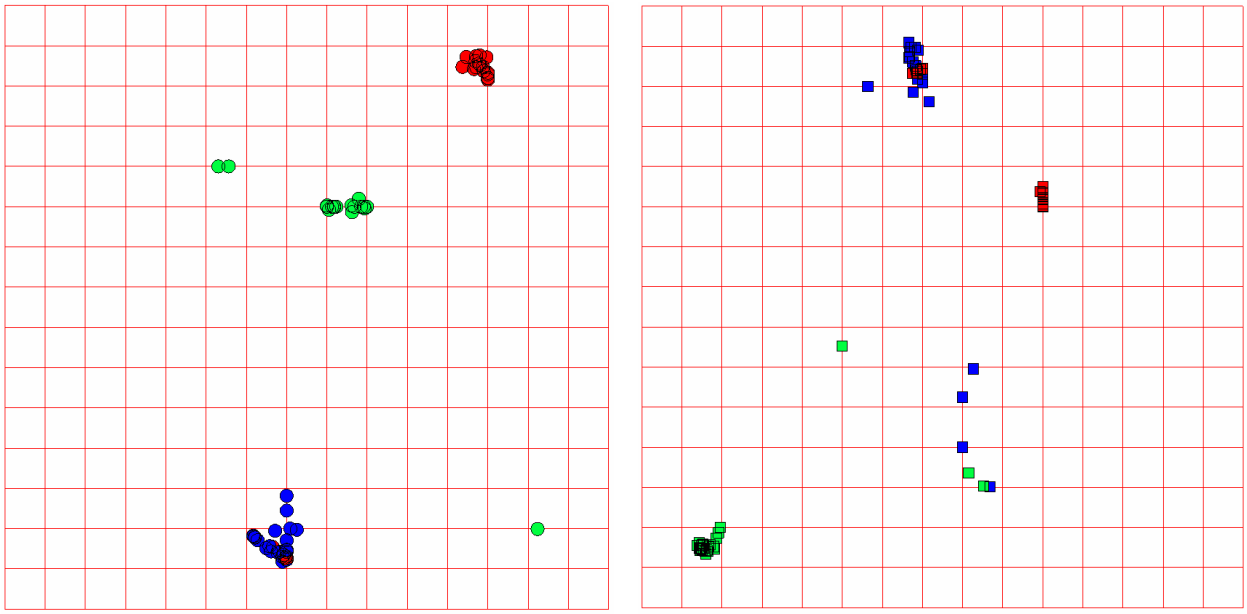


Рисунок 7 — Распределение по синонимичным антикодонам: лейцин. Антикодоны: UAG — красным, CAA — зеленым, UAA– синим. Шаг 1 слева (круги) и шаг 3 справа (квадраты)

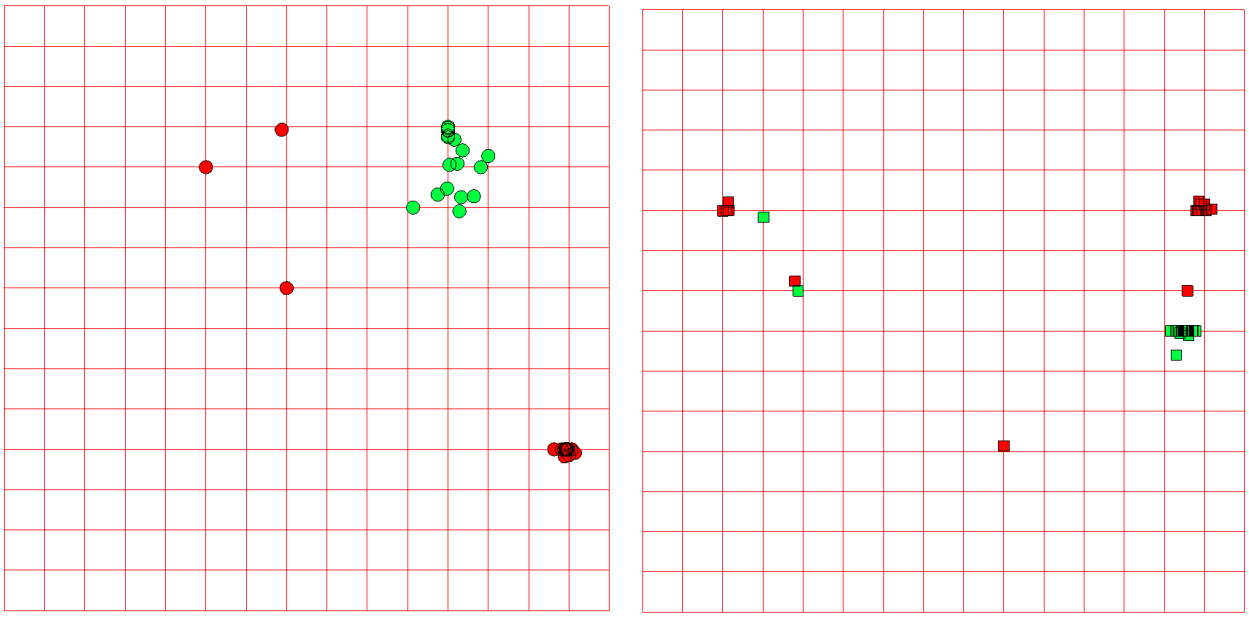


Рисунок 8 — Распределение по синонимичным антикодонам: пролин. Антикодоны: UGG — красным, GGG — зеленым. Шаг 1 слева (круги) и шаг 3 справа (квадраты)

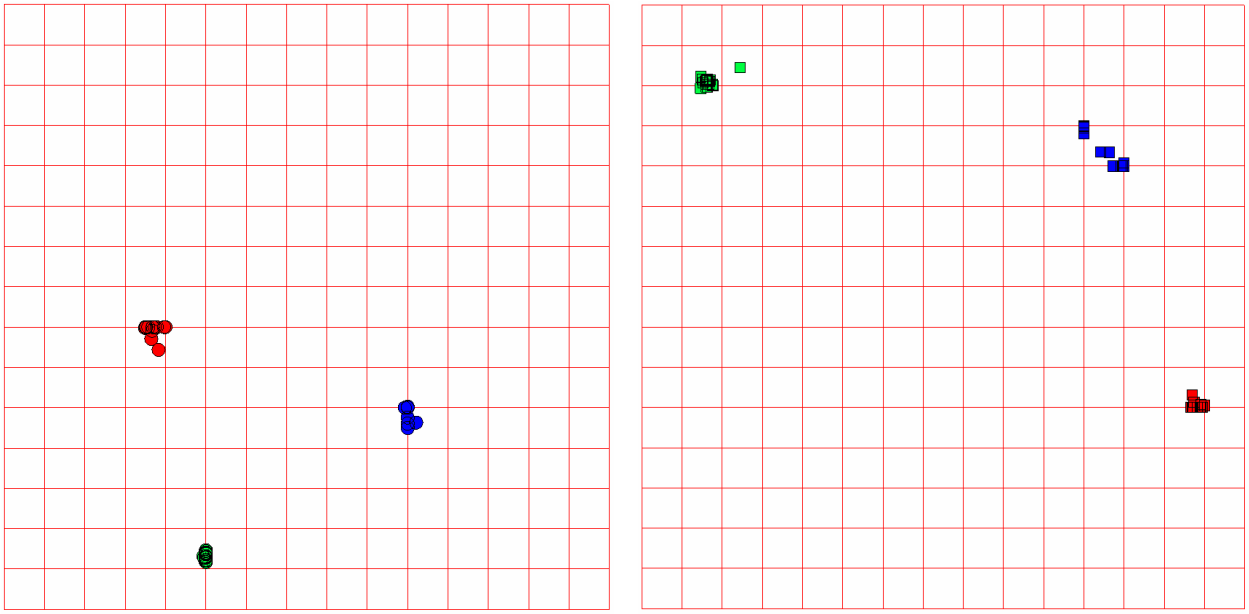


Рисунок 9 — Распределение по синонимичным антикодонам: аргинин. Антикодоны: ACG — красным, UCU — зеленым, CCG– синим. Шаг 1 слева (круги) и шаг 3 справа (квадраты)

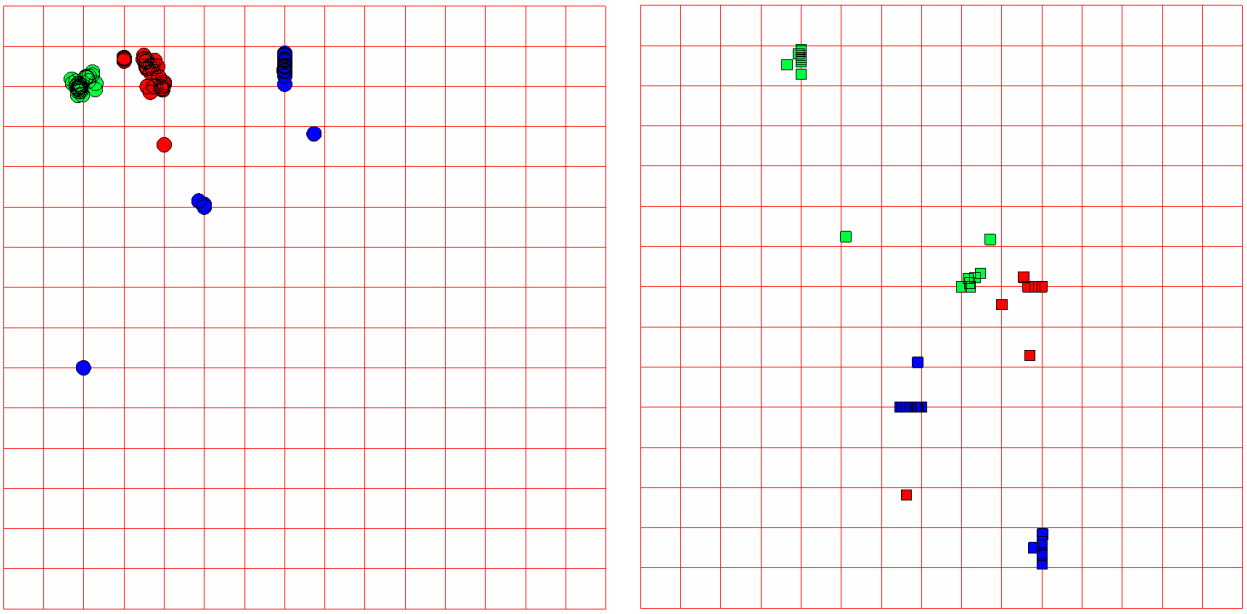


Рисунок 10 — Распределение по синонимичным антикодонам: серин. Антикодоны: GCU — красным, UGA — зеленым, GGA– синим. Шаг 1 слева (круги) и шаг 3 справа (квадраты)

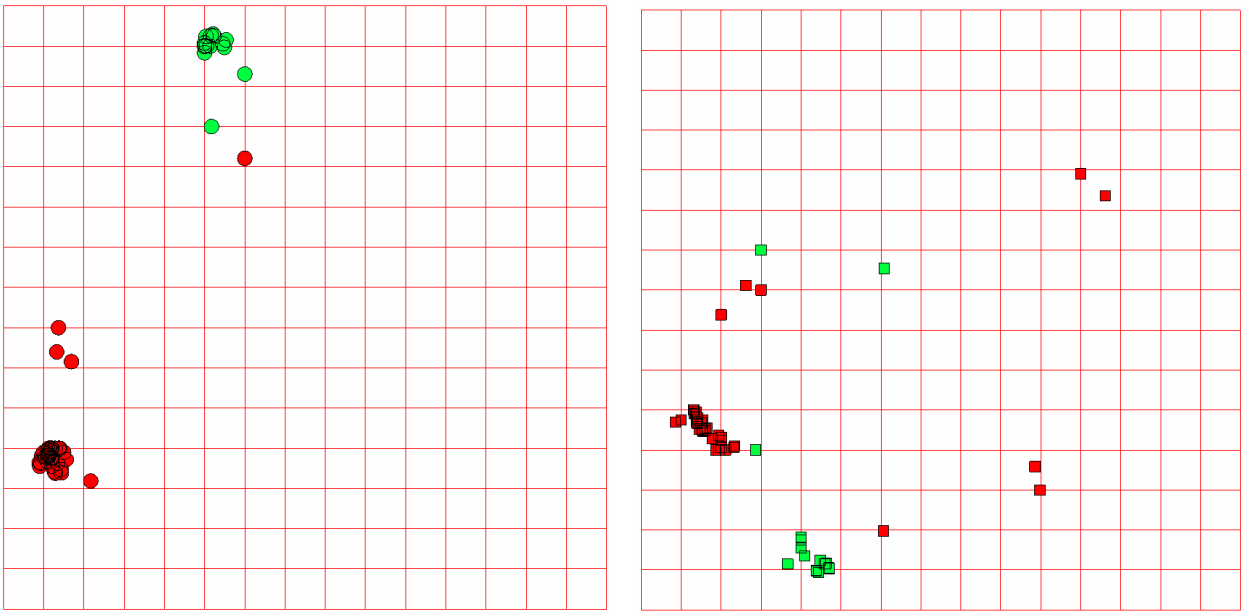


Рисунок 11 — Распределение по синонимичным антикодонам: треонин. Антикодоны: GGU — красным, UGU — зеленым. Шаг 1 слева (круги) и шаг 3 справа (квадраты)

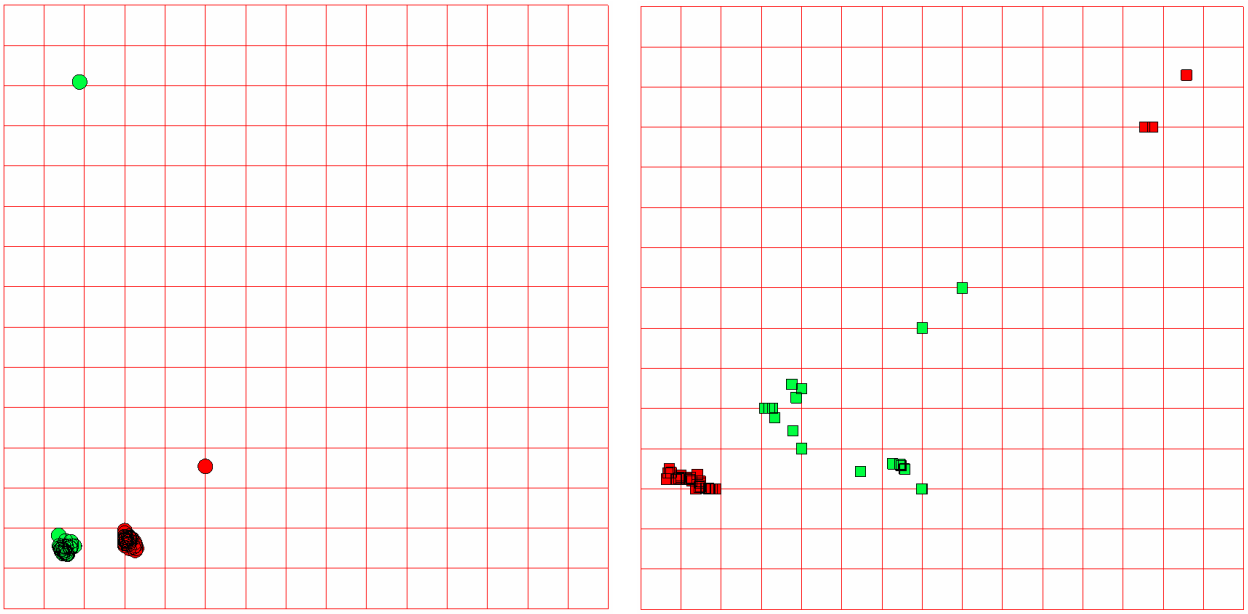


Рисунок 12 — Распределение по синонимичным антикодонам: валин. Антикодоны: GAC — красным, UAC — зеленым. Шаг 1 слева (круги) и шаг 3 справа (квадраты)

5 Обсуждение

5.1 Краткий обзор полученных результатов

Изучению структуры и свойств биологических объектов посвящено большое количество работ. Вид генетического материала влияет на результаты данного исследования существенным образом. В работах [23, 24, 30] исследовалась связь между триплетным составом геномов таксономией носителя для хлоропластов и митохондрий. Так как функция у всех хлоропластов (митохондрий соответственно) одна и та же, то выбор генетического материала позволяет редуцировать задачу. В работах [12, 13] так же рассматривалась лишь связь структуры и таксономии на генетическом материале последовательностей зрелых РНК генов 16S РНК бактерий.

Частотные словари и методы кластеризации применялись в исследовании генетического материала организмов и ранее. Например, в работе [31] в качестве структуры рассматривался частотный словарь триплетов, кластеризация проводилась методом динамических ядер. В работе [32] использовались частотные словари, но визуализация и последующий анализ проводились с помощью метода главных компонент.

Результаты данной работы были сравнены с другими данными. В исследовании [27] рассматривали эволюционнные особенности тРНК хлоропластов 12 видов голосеменных, по каждому виду из семейства.

- Подтверждено, что tRNAIle кодирует антикодон CAU, который обычно кодируется tRNAMet. Возможный механизм, который регулирует специфичность этой аминокислоты, может включать модификацию положения неоднозначности в антикодоне с помощью фермента, модифицирующего тРНК.
- Наиболее представленные количественно во всей базе тРНК: серин, лейцин и аргинин.
- Геномы хлоропластов голосеменных кодируют около 30 антикодон-специфических тРНК. В настоящей работе показано, что часто (> 55) используются 32 антикодона, 16 редко (<5).

- По результатам филогении tRNAMet (CAU), tRNAThr (UGU, GGU), tRNAVal (UAC, GAC), tRNAAla (UGC), tRNAPhe (GAA), tRNAArg (UCU), tRNAHis (GUG), tRNAGln (UUG), tRNACys (GCA), tRNALys (UUU), tRNAGlu (UUC), tRNALeu (CAA), tRNAGly (UCC), tRNASer (CGA), tRNAGly (GCC), и tRNAIle (CAU, UAU), как правило, были самыми основными тРНК, и они претерпели дупликацию и диверсификацию генов с образованием других молекул тРНК.
- Гены тРНК хлоропластов голосеменных растений характеризуются предпочтением кодонов.

5.2 Возможная связь между кластеризацией генов тРНК по частотам триплетов и биохимическими свойствами соответствующих им аминокислот

Если гены тРНК делятся на кластеры в связи с функциями этих генов, то естественно появляется вопрос, по какому признаку сгруппированы функции данных генов и кодируемых аминокислот. Распределение генов в пространстве частот было проанализировано по биохимическим, физическим свойствам аминокислот и свойствам основных синтетаз тРНК.

Транспортные РНК, как и аминокислоты, различаются, так что исследователи группируют их в различные классы. Одной из задач работы была проверка того что точки не делятся на кластеры по таксоном признаку. Результаты работы показывают, что это предположение верно. Однако, при анализе распределения генов с синонимичными антикодонами была выявлена следующая характерная картина: если в таком распределении наблюдаются гены, которые не попали ни в один из кластеров, определяемым антикодоном, то, как правило, это гены одного и того же вида. Например, для валина шага 1 два вида не попали ни в кластер по антикодону GAC, ни по антикодону UAC.

Оба вида принадлежат одному древнему семейству *Саговниковые*: *Cycas revoluta* (NC _020319.1), *Cycas panzhihuaensis* (NC _031413.1).

Аминокислоты делятся по путям биосинтеза на несколько семейств:

- Семейство аспартата: аспартат, аспарагин, треонин, изолейцин, метионин, лизин.
- Семейство глутамата: глутамат, глутамин, аргинин, пролин.
- Семейство пирувата: аланин, валин, лейцин.
- Семейство серина: серин, цистеин, глицин.
- Семейство пентоз: гистидин, фенилаланин, тирозин, триптофан.

Распределение точек для шага 1 по путям биосинтеза аминокислот показано на Рисунке 13 (слева).

По составу радикала аминокислот, по заряду и полярности, наличию ароматических групп, содержанию серы не было выявлено связи с помощью кластеризации.

Аминокислоты имеют средний вес в 100 Да, который варьирует от 89 до 204 Да. Распределение по молекулярному весу показано на Рисунке 13 (справа), где зеленым указаны легкие менее 130 Да, красным — тяжелые, более 130 Да.

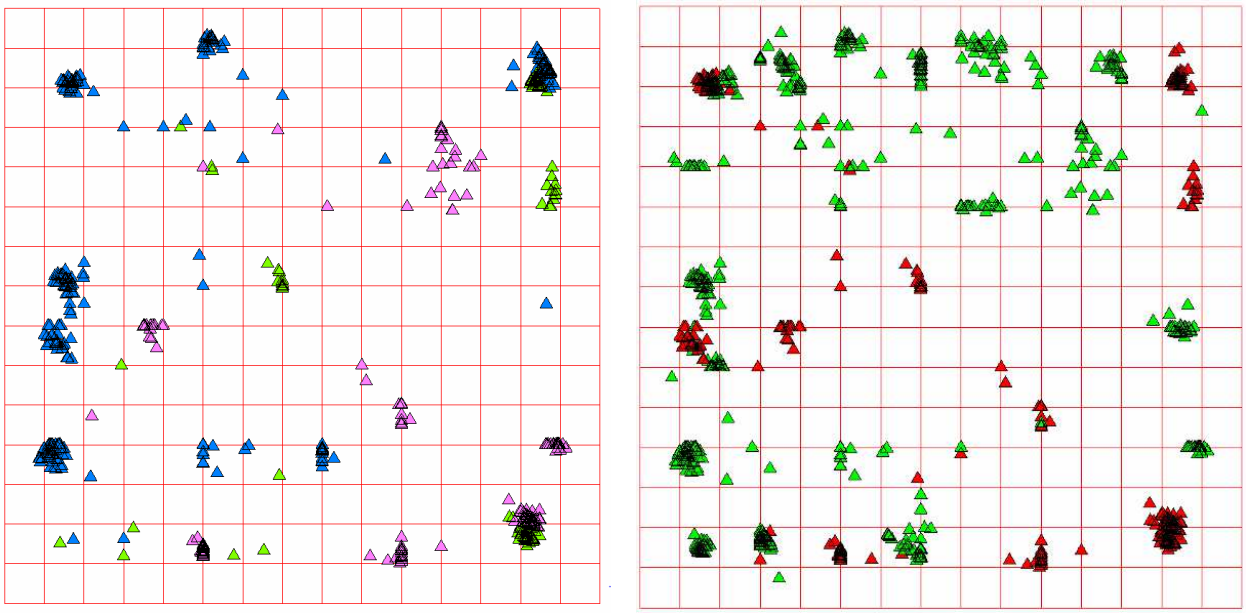


Рисунок 13 — Распределение генов тРНК шага 1 по свойствам аминокислот. Слева — по путям биосинтеза аминокислот. Справа — по молекулярному весу аминокислот.

Сами тРНК принято подразделять по аминоацил-тРНК-синтетазам

на два класса:

- Класс I: валин, изолейцин, лейцин, цистеин, метионин, глутамат, глутамин, аргинин, тирозин, триптофан.
- Класс II: глицин, аланин, пролин, серин, треонин, аспартат, аспарагин, гистидин, фенилаланин.

При этом для лизина существуют синтетазы обоих классов. Этот анализ показал, что такой связи нет (Рисунок 14).

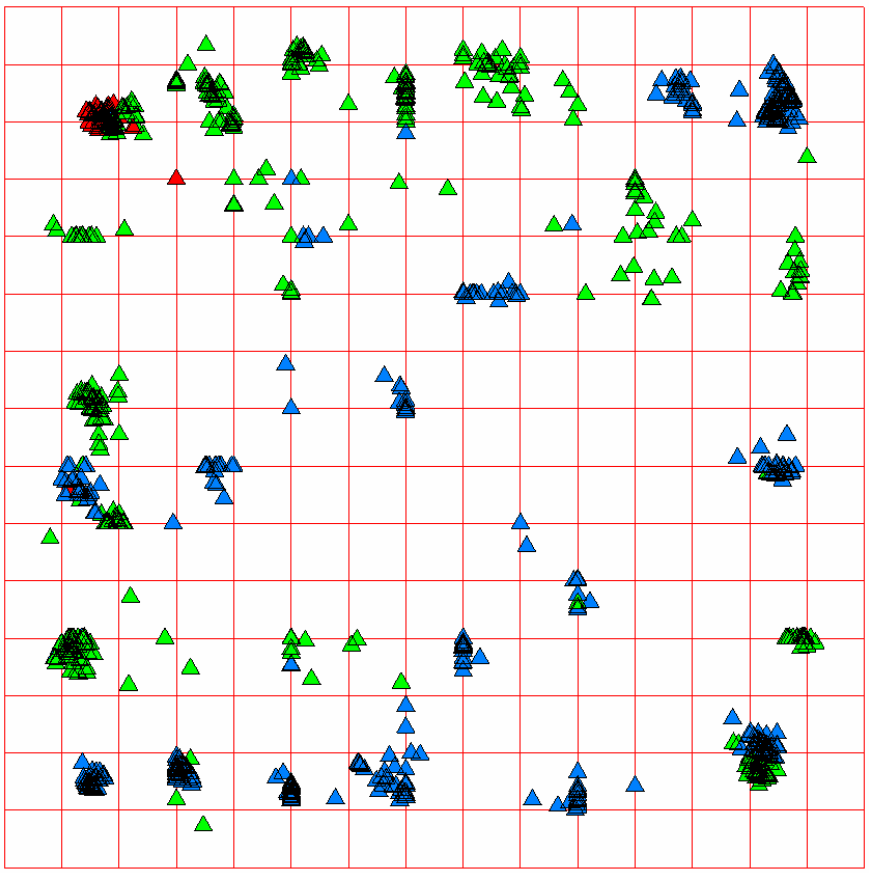


Рисунок 14 — Распределение генов тРНК шага 1 по классам аминоацил-тРНК-синтетаз.

Несмотря на то, что не выявлено яркой зависимости кластеризации с группами по свойствам аминокислот и классам аминоацил-тРНК-синтетаз, гены кластеризуются по синонимам антикодонов для аминокислот. Данное наблюдение может говорить о том, что в эволюционном плане для тРНК имеет большее значение конкретный состав и код триплета антикодона, а не переносимая аминокислота, её пути биосинтеза, биохимические и физические свойства.

ЗАКЛЮЧЕНИЕ

Результаты, полученные в ходе выполнения работы позволили сформулировать однозначные утверждения о связи между функцией, структурой и таксономией генов тРНК. Цель работы достигнута. Все поставленные задачи выполнены:

1) Создана база генов тРНК из полногеномных последовательностей и проанализирована;
2) Построены частотные словари данных последовательностей и проведена кластеризация словарей различными методами кластеризации и визуализации;
3) Проанализировано распределение словарей по кластерам с точки зрения функционального и таксономического состава.

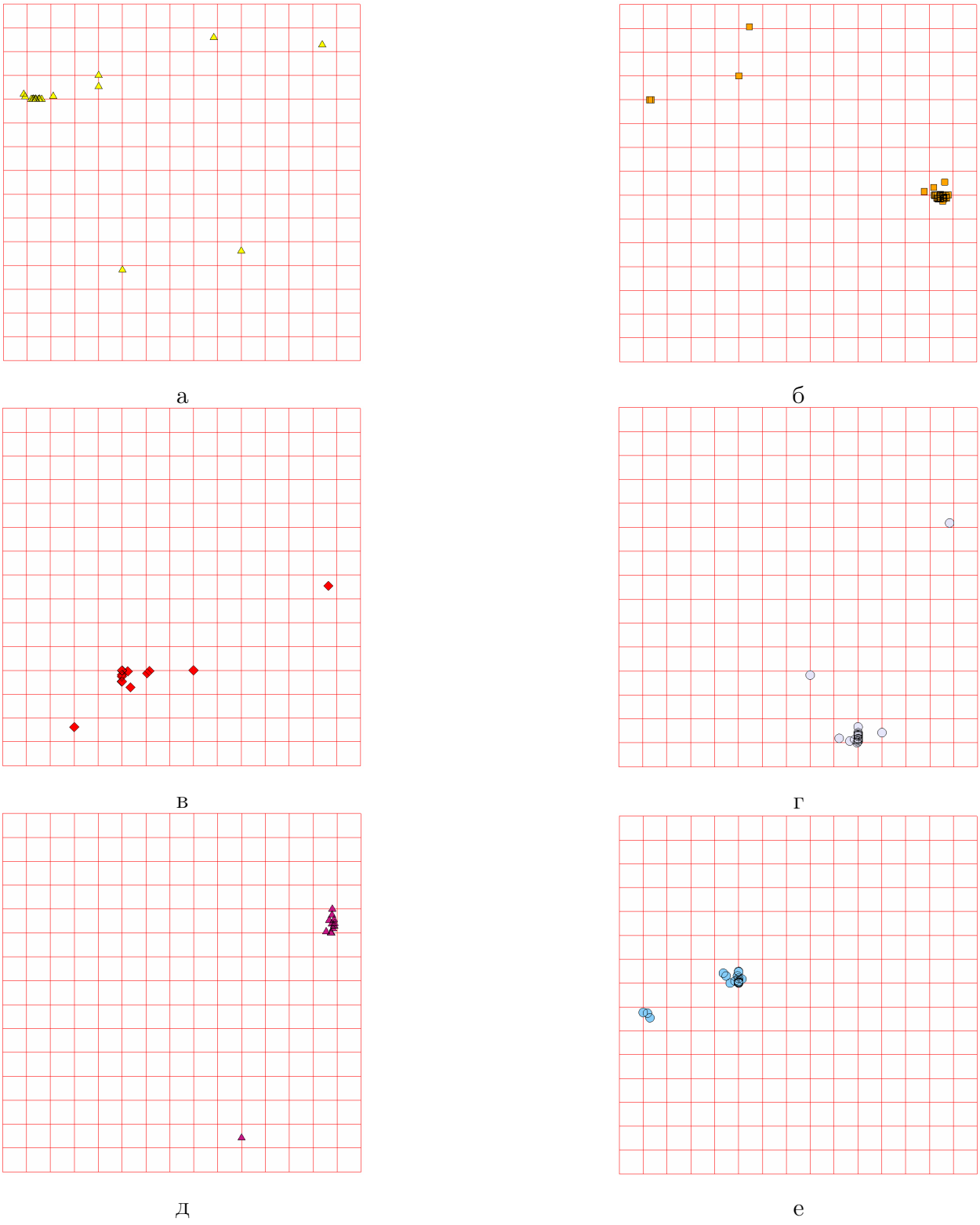
Анализ результатов показал, что для случая генов тРНК хлоропластов голосеменных наблюдается преобладание функции над таксономией. Гены кластеризуются как по кодируемым аминокислотам, так и по синонимам антикодонов для аминокислот. При этом не выявлено однозначной зависимости кластеризации с группами по свойствам аминокислот и классам аминоацил-тРНК-синтетаз. Результаты нашей работы показывают, что для случая генов тРНК не наблюдается никакой связи между видовым составом и составом кластеров, выявленным по частотам триплетов. Точнее, исключения есть, но они малы: если какие-нибудь из генов, кодирующие синонимичные антикодоны для одной аминокислоты не попадали в соответствующие кластеры, определяемые антикодоном, то, как правило, эти гены принадлежат одному и тому же виду. Таким образом, эволюция поддерживает консервацию таких важных генов как тРНК. В базе последовательностей голосеменных было подтверждено, что tRNAIle кодирует антикодон CAU, который обычно кодируется tRNAMet. Подтверждено, что для генов тРНК хлоропластов голосеменных растений характерно предпочтение кодонов.

СПИСОК СОКРАЩЕНИЙ

1. ДНК — дезоксирибонуклеиновая кислота
2. РНК — рибонуклеиновая кислота
3. тРНК — транспортная рибонуклеиновая кислота
4. A — аланин
5. C — цистеин
6. D — Аспартат
7. E — Глутамат
8. F — фенилаланин
9. fM — формилметионин
10. G — глицин
11. H — гистидин
12. I — изолейцин
13. K — лизин
14. L — лейцин
15. M — метионин
16. N — аспарагин
17. P — пролин
18. Q — глутамин
19. R — аргинин
20. S — серин
21. T — треонин
22. v — валин
23. W — триптофан
24. Y — тирозин

ПРИЛОЖЕНИЕ А

Аминокислоты на упругой карте по отдельности



а б в г д е

Рисунок 15 — ШАГ 1 аминокислоты A (а), C (б), D (в), E (г), F (д), fM (е).

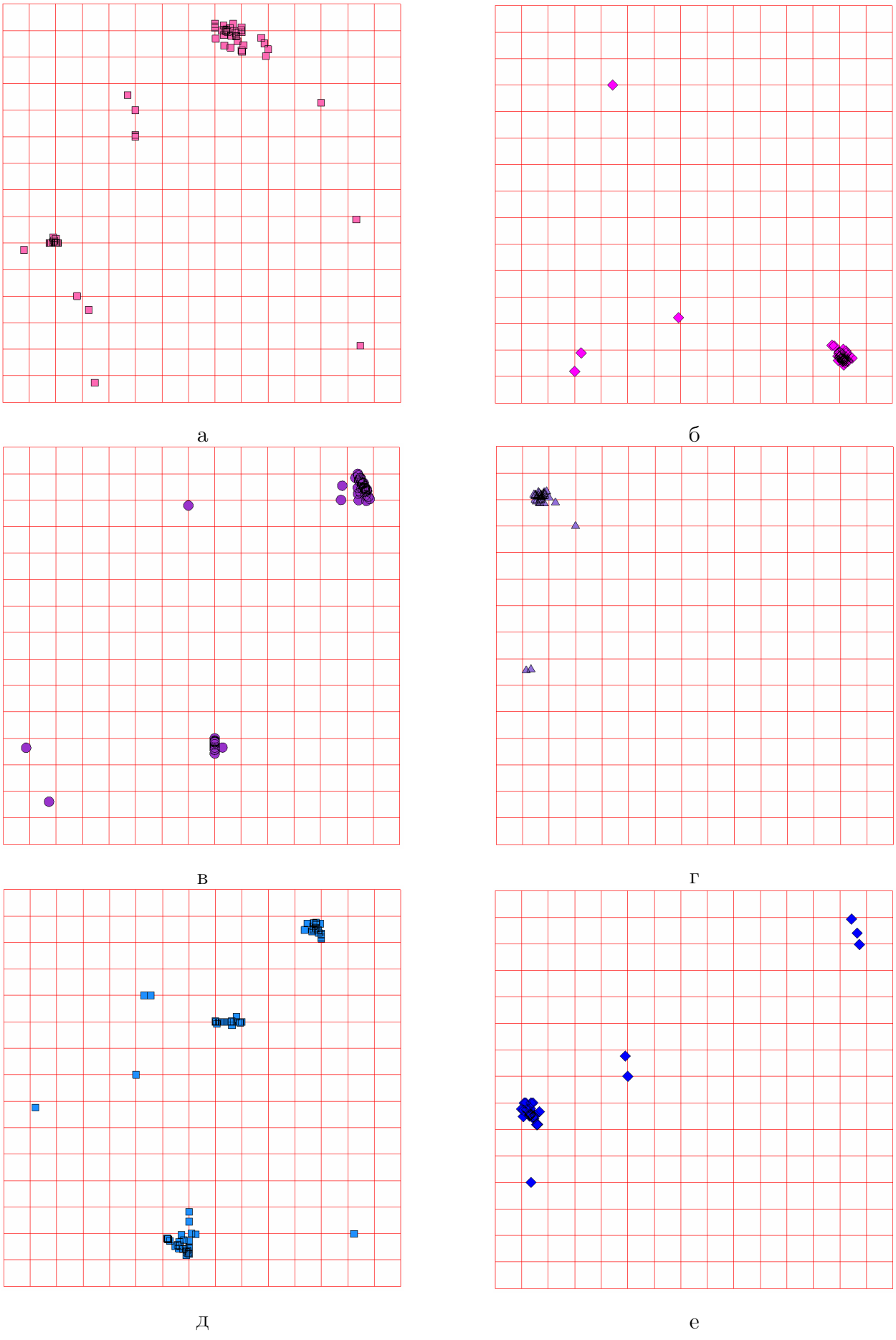


Рисунок 16 — ШАГ 1 аминокислоты G (а), H (б), I (в), K (г), L (д), M (е).

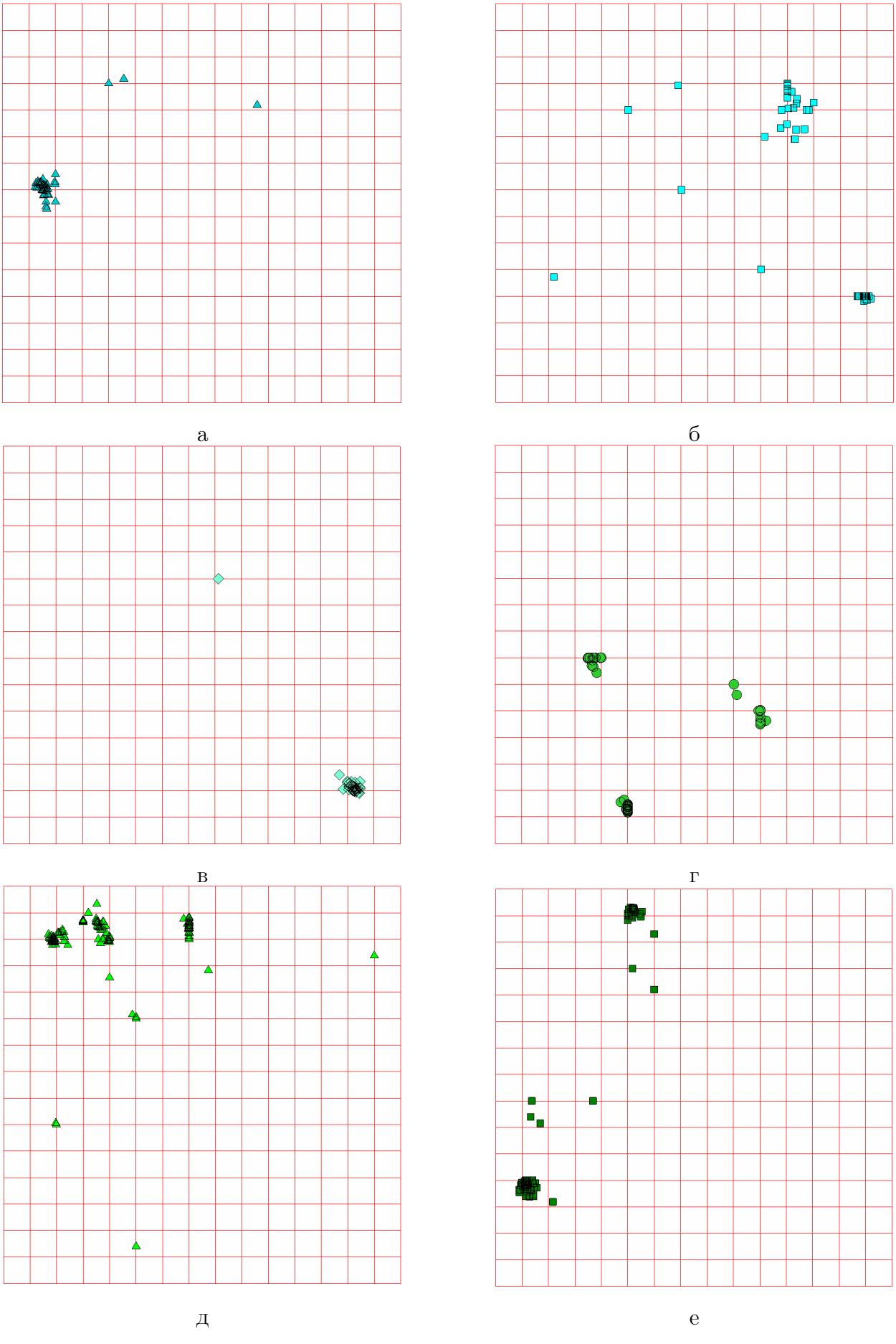
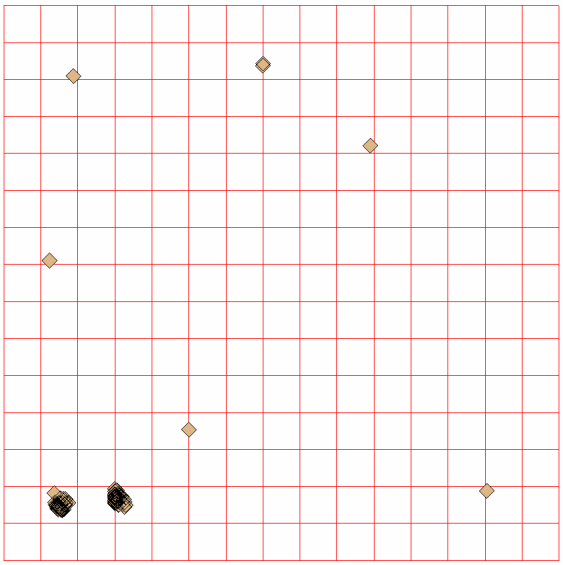
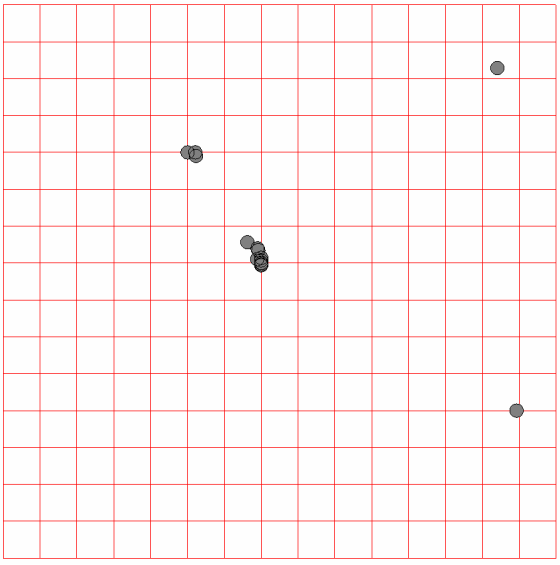


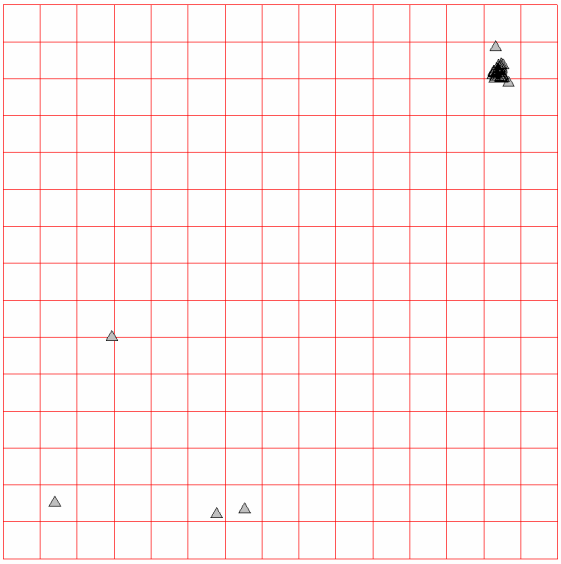
Рисунок 17 — ШАГ 1 аминокислоты N (а), P (б), Q (в), R (г), S (д), T (е).



а

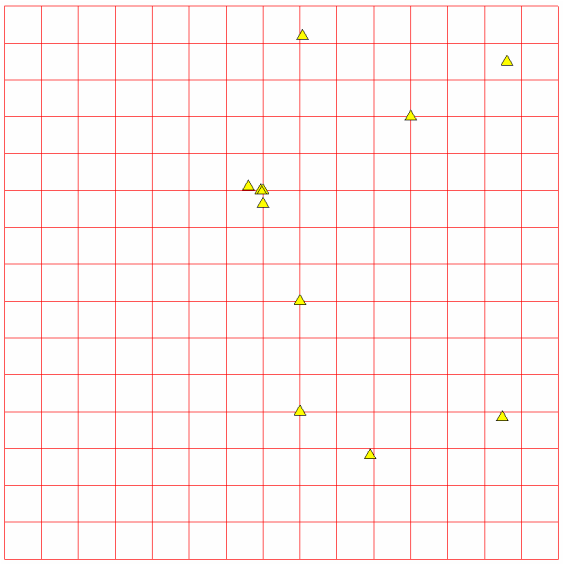


б

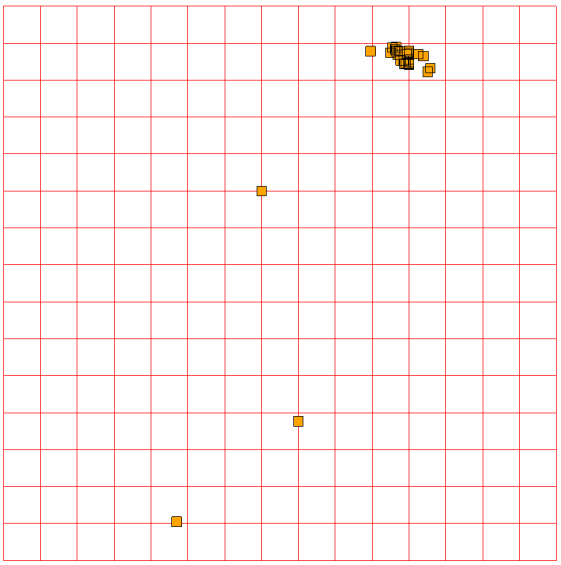


в

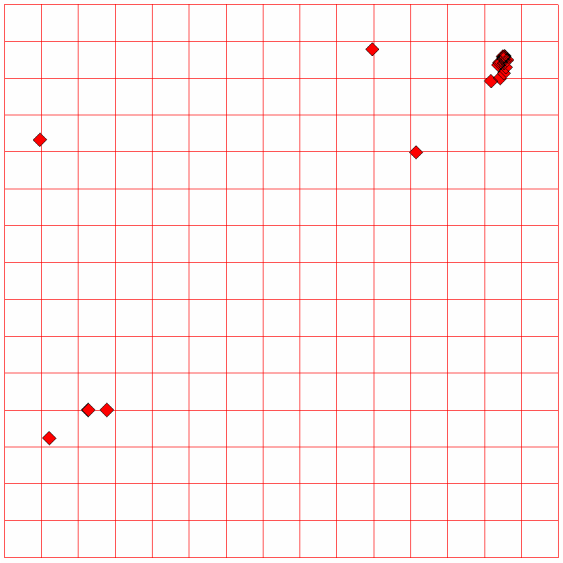
Рисунок 18 — ШАГ 1 аминокислоты V (а), W (б), Y (в).



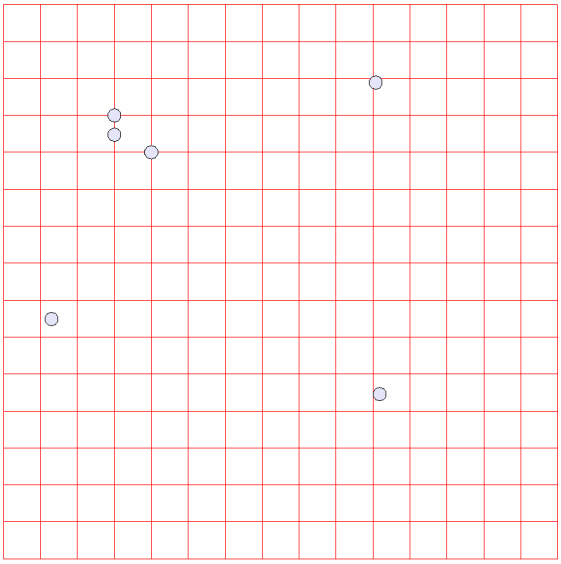
а



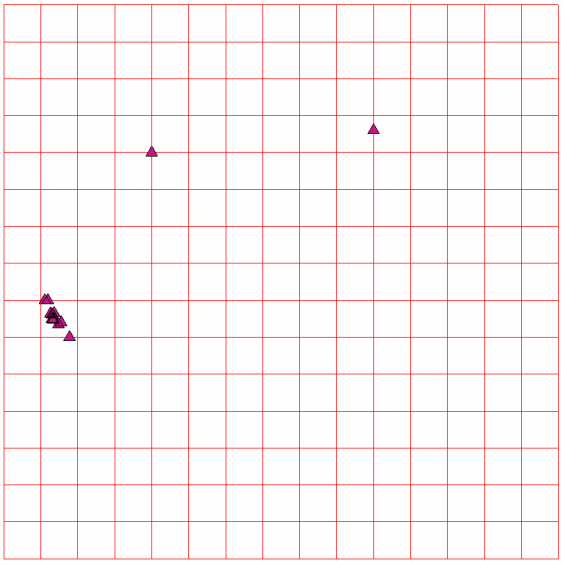
б



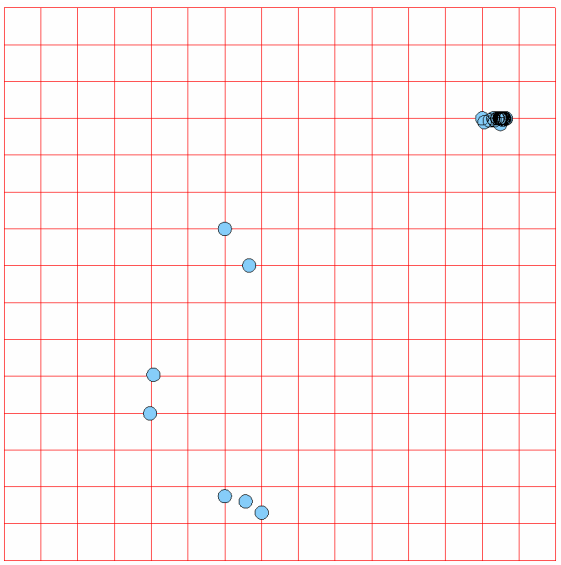
в



г



д



е

Рисунок 19 — ШАГ 3 аминокислоты A (а), C (б), D (в), E (г), F (д), fM (е).

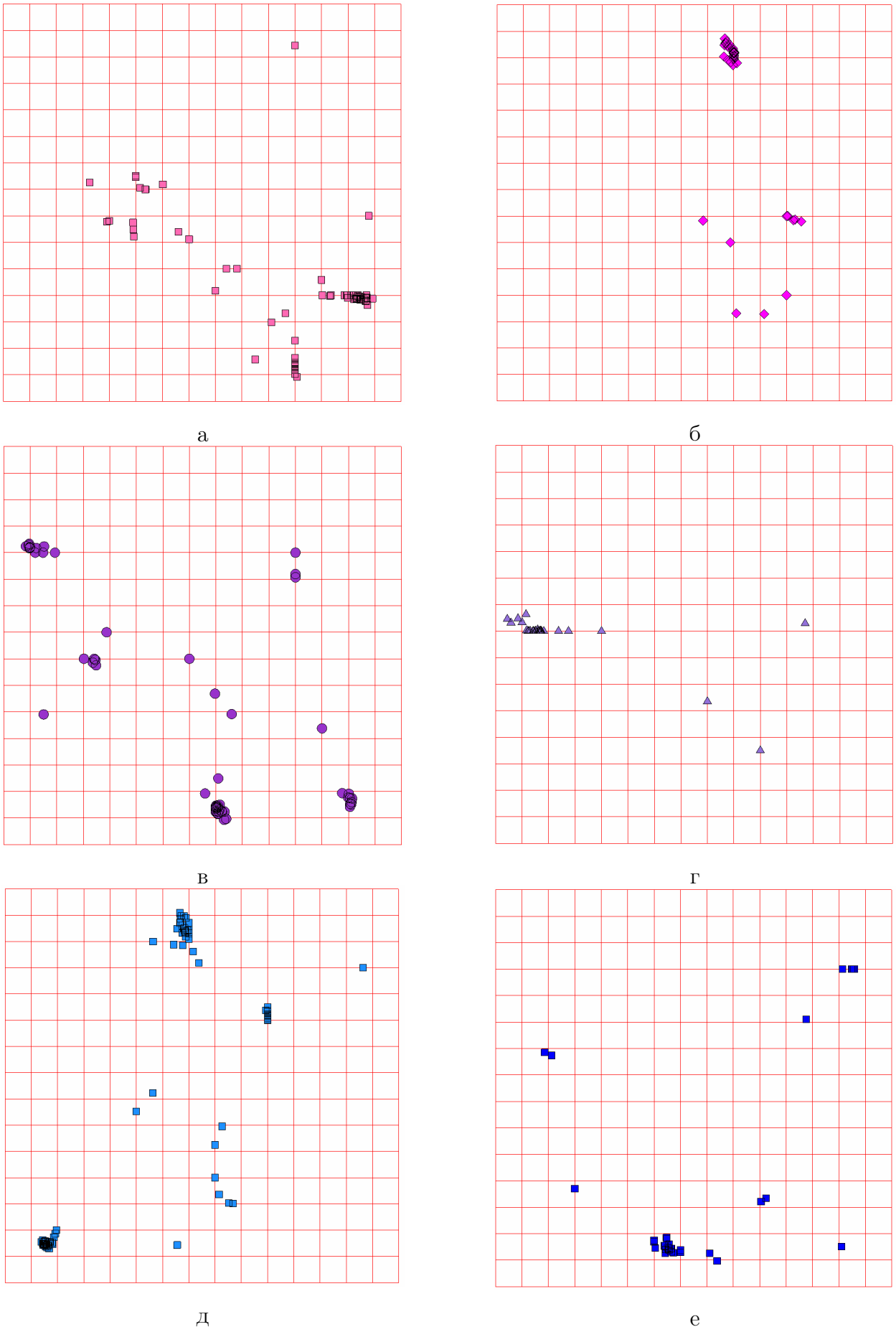
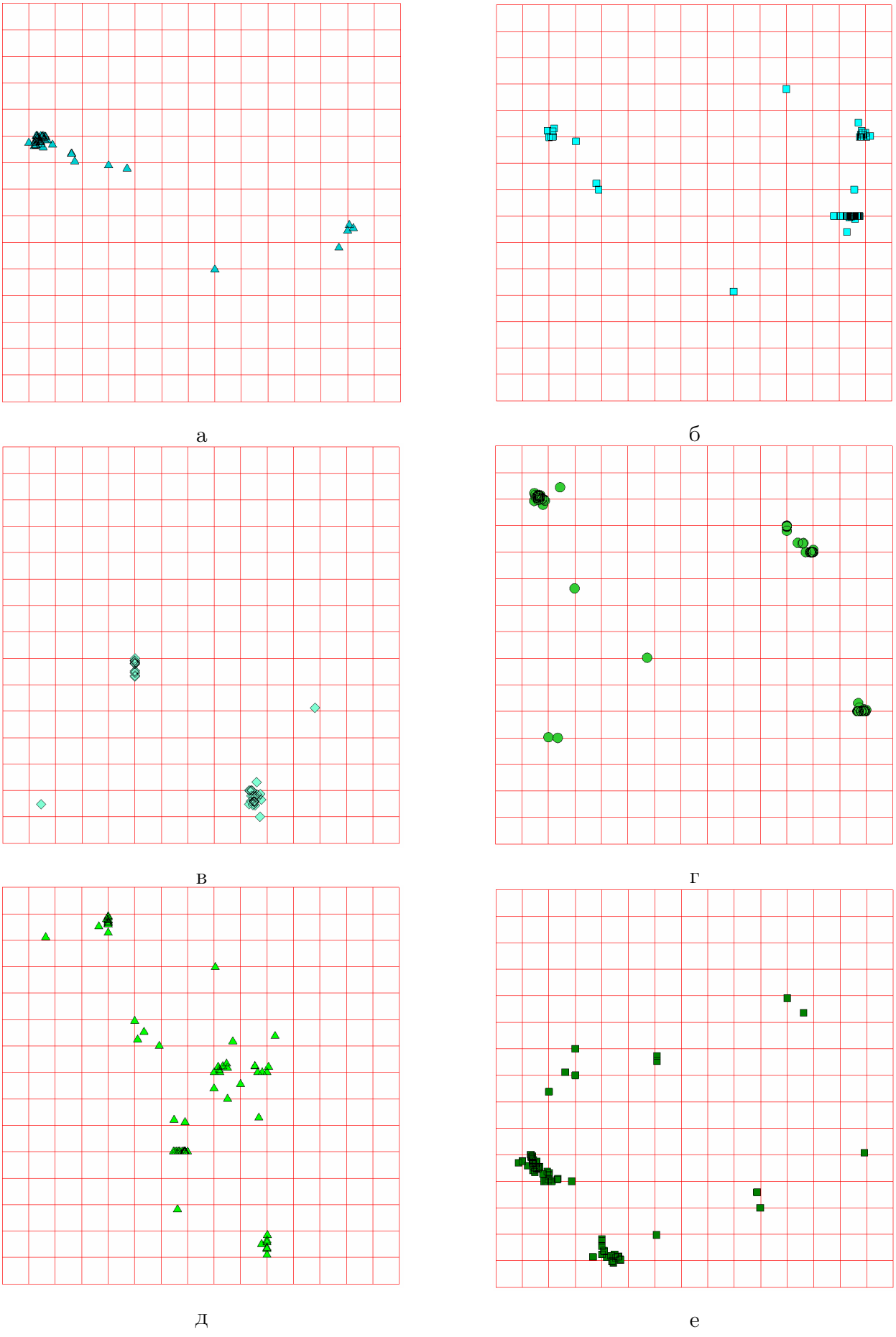
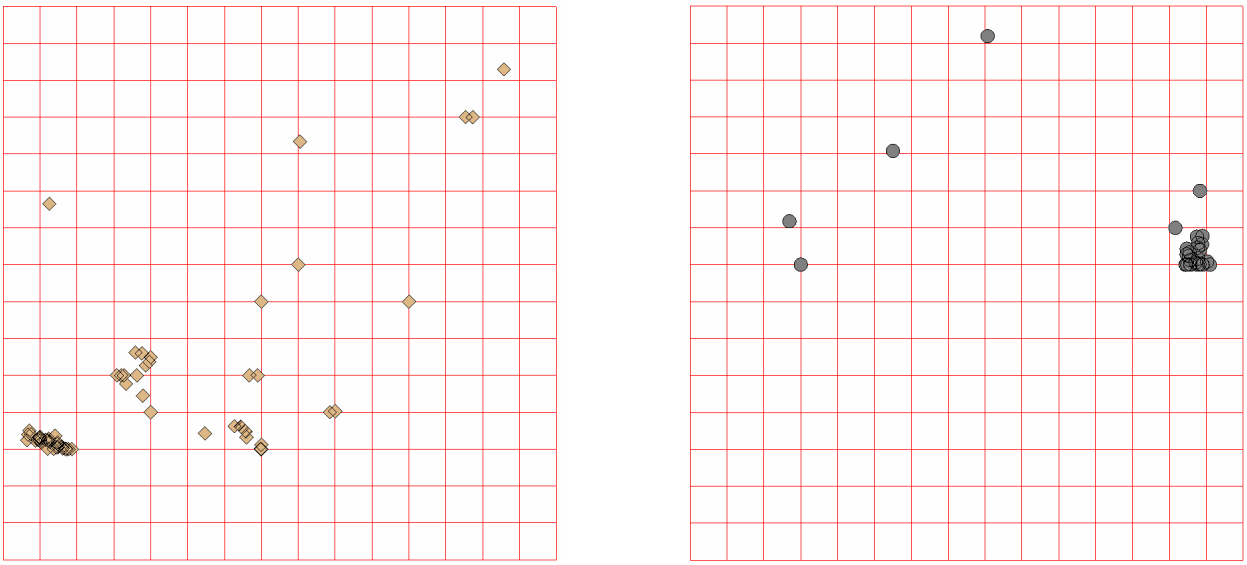


Рисунок 20 — ШАГ 3 аминокислоты G (а), H (б), I (в), K (г), L (д), M (е).

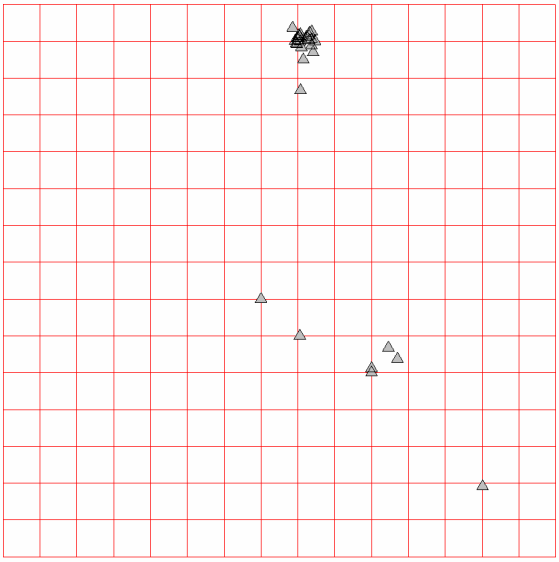


а б в г д е

Рисунок 21 — ШАГ 3 аминокислоты N (а), P (б), Q (в), R (г), S (д), T (е).



а б



в

Рисунок 22 — ШАГ 3 аминокислоты V (а), W (б), Y (в).

ПРИЛОЖЕНИЕ Б

Количество тРНК для каждой аминокислоты из 21. Указаны цвет в кодировке RGB и форма точек на рисунках.

Таблица 5 — Количественное распределение генов тРНК по кодируемым аминокислотам, указаны цвет и форма точек для генов на рисунках

тРНК-аминокислота	N, количество	Цвет в RGB	Символ
A	129	255, 255, 0	△
C	145	255, 165, 0	□
D	148	255, 0, 0	♢
E	143	230, 230, 250	◯
F	143	199, 21, 133	△
G	258	255, 105, 180	□
H	158	255, 0, 255	♢
I	301	153, 50, 204	◯
K	141	147, 112, 219	△
L	419	30, 144, 255	□
N	146	0, 0, 255	♢
fM	142	135,206,250	◯
N	144	0, 206, 209	△
P	275	0, 255, 255	□
Q	166	127, 255, 212	♢
R	371	50, 205, 50	◯
S	427	0, 255, 0	△
T	302	0, 128, 0	□
V	271	222, 184, 135	♢
W	148	128, 128, 128	◯
Y	148	192, 192, 192	△

ПРИЛОЖЕНИЕ В

тРНК по антикодонам, количество которых < 5.

Таблица 6 — Количественное распределение генов тРНК по наименее представленным антикодонам, количество которых <5, указаны цвет и форма точек для генов на рисунках

тРНК-аминокислота	Триплет	N, количество	Цвет в RGB	Символ
A	AGC	2	255, 255, 0	△
A	GGC	2	255, 255, 0	□
C	ACA	2	255, 165, 0	□
D	AUC	2	255, 0, 0	♢
E	TTC	1	230, 230, 250	◯
G	CCC	2	255, 105, 180	□
K	AAA	3	147, 112, 219	△
L	AUG	5	30, 144, 255	□
P	CGG	2	0, 255, 255	□
R	CCU	1	50, 205, 50	◯
S	AGA	2	0, 255, 0	△
S	CGA	1	0, 255, 0	♢
T	CGU	2	0, 128, 0	□
V	CAC	4	222, 184, 135	♢
V	AAC	2	222, 184, 135	◯
Y	AUA	2	192, 192, 192	△

ПРИЛОЖЕНИЕ Г

Количество точек по аминокислотам, гены, у которых есть синонимы (последний столбец — цвет точек на картинках кластеризации по синонимам).

Таблица 7 — Количественное распределение генов тРНК по синонимам антикодонов, указан цвет точек для генов на рисунках

тРНК-аминокислота	Кодон	N, количество	Цвет
G	UCC	114	red
G	GCC	87	green
I	CAU	123	red
I	GAU	107	green
L	UAG	119	red
L	CAA	113	green
L	UAA	98	blue
P	UGG	113	red
P	GGG	102	green
R	ACG	121	red
R	UCU	115	green
R	CCG	55	blue
S	GCU	118	red
S	UGA	113	green
S	GGA	99	blue
T	GGU	132	red
T	UGU	106	green
V	GAC	109	red
V	UAC	105	green

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

[1] AE Bondarev. Visual analysis and processing of clusters structures in multidimensional datasets. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42(2/W4), 2017.

[2] Jeremy Burgess. *Introduction to plant cell development.* CUP Archive, 1985.

[3] N.A. Campbell and J.B. Reece. *Biology.* Global Edition. Pearson, 2011.

[4] Maarten JM Christenhusz, James L Reveal, Aljos Farjon, Martin F Gardner, Robert R Mill, and Mark W Chase. A new classification and linear sequence of extant gymnosperms. *Phytotaxa*, 19(1):55–70, 2011.

[5] Peter Civáň, Peter G Foster, Martin T Embley, Ana Seneca, and Cymon J Cox. Analyses of charophyte chloroplast genomes help characterize the ancestral chloroplast genome of land plants. *Genome Biology and Evolution*, 6(4):897–911, 2014.

[6] Guy Drouin, Hanane Daoud, and Junnan Xia. Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. *Molecular Phylogenetics and Evolution*, 49(3):827–831, 2008.

[7] D. Duchêne and L. Bromham. Rates of molecular evolution and diversification in plants: chloroplast substitution rates correlate with species-richness in the proteaceae. *BMC Evolutionary Biology*, 13:65 – 65, 2012.

[8] K. Fukunaga. *Introduction to statistical pattern recognition.* Academic Press, London, 1990.

[9] A. N. Gorban and A. Yu. Zinovyev. Principal manifolds for data visualisation and dimension reduction. In A N Gorban, B Kégl, D Wünsch, and A Yu Zinovyev, editors, *Lecture Notes in Computational Science and*

Engineering, volume 58, pages 153–176. Springer, Berlin – Heidelberg – New York, 2nd edition, 2007.

[10] Alexander N. Gorban and Andrei Zinovyev. Principal manifolds and graphs in practice: From molecular biology to dynamical systems. *International Journal of Neural Systems*, 20(03):219–232, 2010. PMID: 20556849.

[11] Alexander N. Gorban and Andrei Yu. Zinovyev. Fast and user-friendly non-linear principal manifold learning by method of elastic maps. In *2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2015, Campus des Cordeliers, Paris, France, October 19-21, 2015*, pages 1–9, 2015.

[12] AN Gorban, TG Popova, and MG Sadovsky. Classification of symbol sequences over their frequency dictionaries: towards the connection between structure and natural taxonomy. *Open Systems & Information Dynamics*, 7(1):1–17, 2000.

[13] A.N. Gorban, T.G. Popova, M.G. Sadovsky, and D.C. Wunsch. Information content of the frequency dictionaries, reconstruction, transformation and classification of dictionaries and genetic texts. In *Intelligent Engineering Systems Through Artificial Neural Networks*, pages 657–663. American Society of Mechanical Engineers (ASME), 2001.

[14] Manolo Gouy and Christian Gautier. Codon usage in bacteria: correlation with gene expressivity. *Nucleic acids research*, 10(22):7055–7074, 1982.

[15] Ruth Hershberg and Dmitri A Petrov. Selection on codon bias. *Annual review of genetics*, 42:287–299, 2008.

[16] Anita K Hopper and Eric M Phizicky. trna transfers to the limelight. *Genes & development*, 17(2):162–180, 2003.

[17] Chun Y Huang, Michael A Ayliffe, and Jeremy N Timmis. Direct measurement of the transfer rate of chloroplast dna into the nucleus. *Nature*, 422(6927):72–76, 2003.

[18] Shiv Shankhar Kaundun and Satoru Matsumoto. Molecular evidence for maternal inheritance of the chloroplast genome in tea, camellia sinensis (l.) o. kuntze. *Journal of the Science of Food and Agriculture*, 91(14):2660–2663, 2011.

[19] Sebastian Kirchner and Zoya Ignatova. Emerging roles of trna in adaptive translation, signalling dynamics and disease. *Nature Reviews Genetics*, 16(2):98–112, 2015.

[20] Shai Koussevitzky, Ajit Nott, Todd C Mockler, Fangxin Hong, Gilberto Sachetto-Martins, Marci Surpin, Jason Lim, Ron Mittler, and Joanne Chory. Signals from chloroplasts converge to regulate nuclear gene expression. *Science*, 316(5825):715–719, 2007.

[21] Geoffrey I. McFadden. Chloroplast Origin and Integration1. *Plant Physiology*, 125(1):50–53, 01 2001.

[22] Tapan K Mohanta, Asad S Syed, Fuad Ameen, and Hanhong Bae. Novel genomic and evolutionary perspective of cyanobacterial trnas. *Frontiers in genetics*, 8:200, 2017.

[23] Michael Sadovsky, Yulia Putintseva, Anna Chernyshova, and Vaselina Fedotova. Genome structure of organelles strongly relates to taxonomy of bearers. In *International Conference on Bioinformatics and Biomedical Engineering*, pages 481–490. Springer, 2015.

[24] Michael G Sadovsky, Natalia A Zaitseva, and Yulia A Putintseva. System biology on mitochondrion genomes. In *The Third International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies*, pages 61–66, 2011.

[25] Xiao-Quan Wang and Jin-Hua Ran. Evolution and biogeography of gymnosperms. *Molecular Phylogenetics and Evolution*, 75:24–40, 2014.

[26] Jeremy E Wilusz. Controlling translation via modulation of trna levels. *Wiley Interdisciplinary Reviews: RNA*, 6(4):453–470, 2015.

[27] Ting-Ting Zhang, Yi-Kun Hou, Ting Yang, Shu-Ya Zhang, Ming Yue, Jianni Liu, and Zhonghu Li. Evolutionary analysis of chloroplast trna of gymnosperm revealed the novel structural variation and evolutionary aspect. *PeerJ*, 8:e10312, 2020.

[28] Б Альбертс, Д Брей, К Хопкин, А Джонсон, Дж Льюис, М Рэфф, К Робертс, and П Уолтер. *Основы молекулярной биологии клетки*. М.: Бином, 2015.

[29] Н. Н. Бугаенко, А. Н. Горбань, and М. Г. Садовский. Информационная ёмкость нуклеотидных последовательностей и их фрагментов. *Биофизика*, 42(5):1047–1053, 1997.

[30] Александр Н. Горбань, Т. Г. Попова, and М. Г. Садовский. Классификация нуклеотидных последовательностей по частотным словарям обнаруживает связь между их структурой и таксономическим положением организмов. *Журнал общей биологии*, 64(1):65–77, 2003.

[31] М.Г. Садовский and А.И. Чернышова. Выявление связи структуры и таксономии геномов хлоропластов методом динамических ядер. *Фундаментальные исследования*, 3(11), 2014.

[32] М.Ю. Сенашова and М.Г. Садовский. Семикластерная структура геномов хлоропластов отражает филогению их носителей. *Международный журнал прикладных и фундаментальных исследований*, 12(7):1167–1173, 2016.

Федеральное государственное автономное образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
Институт фундаментальной биологии и биотехнологии
Кафедра биофизики

УТВЕРЖДАЮ:
заведующий кафедрой

В. А. Кратасюк
"28" июня 2021 г.

БАКАЛАВРСКАЯ РАБОТА

03.03.02 Физика

СВЯЗЬ СТРУКТУРЫ И ФУНКЦИИ ГЕНОВ ТРАНСПОРТНЫХ РНК ХЛОРОПЛАСТОВ ГОЛОСЕМЕННЫХ РАСТЕНИЙ

Руководитель: _____ д.ф.-м.н., проф. М. Г. Садовский
дата, подпись 28.06.2021 уч.степень, должность

Выпускник: _____ Т. О. Шпагина
дата, подпись 28.06.2021

Красноярск 2021