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

Институт фундаментальной биологии и биотехнологии
Кафедра биофизики

УТВЕРЖДАЮ
Заведующий кафедрой
_____ В. А. Кратасюк
« _____» _____ 2016 г.

БАКАЛАВРСКАЯ РАБОТА

06.03.01 Биология

СБОРКА, АННОТИРОВАНИЕ И АНАЛИЗ СТАТИСТИЧЕСКИХ И КОМБИНАТОРНЫХ СВОЙСТВ ХЛОРОПЛАСТНОГО ГЕНОМА ЛИСТВЕННИЦЫ СИБИРСКОЙ

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

Выпускник _____ Е. И. Бондар

Красноярск 2016

РЕФЕРАТ

Выпускная квалификационная работа содержит 48 страниц текстового документа, 1 приложение, 30 использованных источников, 8 рисунков, 3 таблицы, 11 формул.

ЛИСТВЕННИЦА СИБИРСКАЯ, ХЛОРОПЛАСТНЫЙ ГЕНОМ, АННОТИРОВАНИЕ ГЕНОМА, КЛАСТЕРИЗАЦИЯ, УПРУГИЕ КАРТЫ

Цель исследования – сборка и аннотирование хлоропластного генома лиственницы сибирской (*Larix sibirica* Ledeb., 1833), поиск однонуклеотидных полиморфизмов (SNPs), а также исследование геномной последовательности с помощью математических методов анализа многомерных данных.

Объектом исследования являются свойства генома хлоропласта лиственницы сибирской *Larix sibirica* Ledeb.

Получена полная последовательность хлоропластного генома лиственницы сибирской длиной 122561 п.н., а также его аннотация. Геном содержит 121 кодирующий участок, из которых 34 соответствуют генам РНК и 87 – CDS. Среди генетического материала трех деревьев, произрастающих в разных регионах России найдено 13 SNP. Составлен и проанализирован частотный словарь хлоропластного генома при помощи метода упругих карт. Выявлена семикластерная структура и определён генный состав кластеров.

Оглавление

ВВЕДЕНИЕ

Широкое распространение технологий секвенирования нового поколения в последние десятилетия, позволившее сравнительно быстро и с гораздо меньшими затратами проводить геномные исследования, дало толчок развитию разнообразных методов биоинформатики для целей обработки получаемых данных. Подобные ДНК-технологии находят своё применение, в частности, в области изучения особенностей процессов, происходящих в популяциях, для решения задач селекции, сохранения биоразнообразия.

Хвойные растения составляют огромную часть лесного массива Сибири и Дальнего востока. Для геномов растений характерно наличие строго определенной системы наследования – передача митохондриальной ДНК по материнской линии, а хлоропластной – по отцовской [1]. Работа с митохондриальными геномами хвойных сопряжена с определенными трудностями – он сравнительно велик и обладает высоким уровнем структурных перестроек. Поэтому именно последовательности хлоропластной ДНК разных видов хвойных являются на сегодня важным источником генетических маркеров в популяционных и филогенетических исследованиях.

К настоящему моменту из более чем 100 депонированных в базе данных NCBI геномов хлоропластов семейства *Pinaceae* большинство относятся к роду *Pinus*, и лишь два вида – к роду *Larix*. По состоянию на 20 апреля 2015 года в базе данных Genbank опубликовано 117 хлоропластных геномов представителей семейства Сосновые. Для представителей рода *Larix* опубликованы две последовательности хлоропластных геномов для видов *L.decidua* Mill. и *L.occidentalis* Nutt. Полная последовательность хлоропластного генома *L. sibirica* в базе данных отсутствует, имеется лишь 11 последовательностей некоторых генов и их частей.

Цель настоящей работы — сборка и аннотирование хлоропластного генома лиственницы сибирской (*Larix sibirica* Ledeb., 1833), поиск

однонуклеотидных полиморфизмов (SNPs), а также исследование геномной последовательности с помощью математических методов анализа многомерных данных.

В связи с этим были поставлены следующие задачи:

- Сборка хлоропластного генома *L. Sibirica;*
- Поиск кодирующих участков и их аннотирование;
- Поиск однонуклеотидных полиморфизмов среди генетического материала трех деревьев, произрастающих в разных регионах России.
- Анализ частотного словаря хлоропластного генома *Larix sibirica* в программе VidaExpert.

1 Обзор литературы

1.1 Актуальность исследования геномов хвойных

Хвойный лес – один из интереснейших объектов живой природы. Такие ценные компоненты бореальных лесов России как лиственница и сосна имеют огромное экономическое, экологическое и эстетическое значение, и играют важнейшую биосферную роль в регуляции глобального климата. Ввиду этого разработка научно обоснованных подходов по сохранению и преумножению генетических ресурсов лиственницы и других важных пород хвойных, а также выведение новых ценных древесных пород, устойчивых к засухе и вредителям имеют неоспоримую практическую значимость.

На сегодняшний день геномика – это интеграционная наука, основанная на современных методах секвенирования и анализа ДНК методами биоинформатики для выяснения её организации, функции и регуляции. Без знания структуры и функции генома невозможно понять развитие и эволюцию организмов. К сожалению, полноценное изучение данного направления тормозится практически полным отсутствием данных о геноме и генах хвойных, контролирующих важные адаптивные и селекционные признаки.

При наличии расшифрованного генома, есть возможность определить относительно полный набор генов, а затем через сравнительный анализ выделить участки низкой и высокой изменчивости генов и использовать их в создании специальных маркеров для изучения связи этой генетической изменчивости с изменчивостью сложных адаптивных и хозяйственно-ценных признаков. А это в дальнейшем позволит решать практические задачи селекции.

Получение эталонных, полностью аннотированных геномов основных лесообразующих пород способствует получению важной информации и является, безусловно, необходимым этапом в разработке высокоинформативных молекулярно-генетических маркёров, которые могут

быть эффективно использованы для определения происхождения древесины, изучения и мониторинга генетической изменчивости хвойных лесов, их адаптации к изменению климата, а также для решения задач геномной селекции на высокое качество древесины, скорость роста, прирост биомассы, устойчивость к вредителям, болезням и другие важные селекционные признаки.

Расшифровкой столь больших геномов занимаются по всему миру лишь несколько лабораторий, так как огромные размеры растительных геномов, а в особенности геномов хвойных, являются серьезным препятствием для секвенирования. Современные методы секвенирования генома позволяют относительно легко читать отдельные нуклеотидные фрагменты ДНК, но лишь в коротких последовательностях. В случае хвойных необходимо сопоставлять друг с другом более 10 миллиардов отдельных последовательностей, что является задачей высокой вычислительной сложности.

1.2 Сборка генома

1.2.1 Особенности данных Illumina и оценка качества

1 апреля 1997 года, Паскаль Майер и Лоран Фаринелли представили во Всемирной организации интеллектуальной собственности патент, описывающий новый метод секвенирования ДНК и пробоподготовки методом ПЦР. Данный подход в сочетании с "base-by-base" методом, предложенным британскими химиками Шанкаром Баласубраманианом и Дэвидом Кленерманом, в настоящее время лежит в основе технологии секвенирования Illumina HiSeq.

В отличие от классического «метода обрыва цепи» Сэнгера, Illumina использует так называемую технологию секвенирования посредством синтеза (SBS) – отслеживание присоединения меченых нуклеотидов. 3’-модификация не позволяет ДНК-полимеразе присоединить больше одного нуклеотида. Флуоресценция инициируется коротким импульсом лазера, тип

присоединенного нуклеотида определяется по цвету флуоресцентной метки. Удлинение цепи ДНК при этом происходит постепенно, что позволяет за раз с помощью камеры снимать большие ДНК-чипы. В результате удается определить последовательность ДНК длиной до 250 нуклеотидов. HiSeq 2000 способен секвенировать 6 человеческих геномов за 11 дней. Длина прочтения составляет 100 нуклеотидов, можно получить 600 Gb информации [2].

Одной из основных причин того, что технологии NGS получили такое широкое распространение в последние годы, является высокое качество данных. Тем не менее, не все системы могут предложить тот же уровень качества. Последствия ошибок могут оказаться весьма неприятными: ложноположительные и ложноотрицательные срабатывания могут привести к более высоким затратам и клиническим ошибкам.

Оценка показателя качества – это вероятность того, что нуклеотидное основание было распознано неверно. Высший показатель качества указывает на меньшую вероятность ошибки. Оценка качества по шкале Phred первоначально была разработана программой Phred для помощи в автоматизации ДНК-секвенирования в проекте «Геном человека». Оценки качества (Q-score) Phred назначаются каждому нуклеотидному основанию, определенному в ходе секвенирования [3]. При этом *Q*-score определяется как отрицательный десятичный логарифм вероятности ошибки, возведенной в десятую степень. Так Phred 30 ($Q = -10 \cdot \lg 0{,}001$) соответствует вероятности ошибки $P = 10^{\frac{-30}{10}}$; у Phred 60 ($Q = -10 \cdot \lg 0{,}000001$) вероятность ошибки $P = 10^{\frac{-60}{10}}$.

По сравнению с другими методами секвенирования нового поколения, метод Illumina является наиболее высокопроизводительным. Как следствие, стоимость секвенирования на 1 Gb информации сравнительно невелика. Кроме того, точность при небольшой длине прочтений является достаточно

высокой [4]. Большинство возникающих ошибок секвенирования – неправильное определение присоединенного нуклеотида. Средняя частота ошибок составляет 0,5%, то есть одна ошибка на прочтение в 200 нуклеотидов [2]. Недостатком же технологии является небольшая длина ридов (прочтений) – до 250 нуклеотидов.

Немаловажным аспектом является формат предоставляемых данных. В последнее время стандартом де-факто для хранения данных секвенирования стал текстовый формат FASTQ. Помимо самой нуклеотидной либо аминокислотной последовательности, он включает также соответствующие показатели качества.

1.2.2 Методы ассемблирования

Секвенирование биополимеров (белков и нуклеиновых кислот – ДНК и РНК) – определение их аминокислотной или нуклеотидной последовательности. В результате секвенирования получают формальное описание первичной структуры линейной макромолекулы в виде последовательности мономеров в текстовом виде. Размеры секвенируемых участков ДНК обычно составляют 100-250 пар нуклеотидов при использовании методов высокопроизводительного секвенирования и 1000 пар нуклеотидов при секвенировании по Сэнгеру. В результате секвенирования перекрывающихся участков ДНК получают последовательности участков генов, целых генов, тотальной мРНК и полных геномов организмов.

Сборка генома (ассемблирование) – процесс объединения большого количества коротких фрагментов ДНК (ридов) в одну или несколько длинных последовательностей (контигов и скаффолдов) в целях восстановления последовательностей ДНК хромосом, из которых эти фрагменты были получены в процессе секвенирования. Это необходимо, так как при использовании существующих технологий секвенирования ДНК невозможно прочитать геном целиком, а только лишь небольшие его фрагменты длиной от 20 до 30000 п.н. (в зависимости от используемой технологии).

Сборка генома является очень сложной вычислительной задачей, в частности, потому, что геномы часто содержат большое количество одинаковых повторяющихся последовательностей (так называемые геномные повторы). Эти повторы могут быть длиной в несколько тысяч нуклеотидов, а также встречаться в тысяче различных мест в геноме. Особенно богаты повторами большие геномы растений и животных, в том числе геном человека.

Проблема сборки исходной геномной последовательности может быть проиллюстрирована классическим примером с литературным текстом. Можно представить, что каждый нуклеотид это буква, а триплеты – комбинации из трёх последовательно расположенных нуклеотидов в молекуле нуклеиновой кислоты, образующие кодоны, с помощью которых в информационных рибонуклеиновых кислотах кодируется последовательность расположения аминокислот в белках – это «слова». Из кодонов складываются гены – законченные «предложения», а из совокупности генов формируется «полный» текст, который и называется геномом. Если сравнить геном человека и книгу, например роман Льва Николаевича Толстого «Война и мир», то окажется, что писатель использовал примерно три миллиона букв и знаков препинания, а у человека три миллиарда таких «букв» – нуклеотидов в геноме. То есть каждый человек носит в себе условную тысячу томов известного романа. При секвенировании происходит многократное копирование всего объема текста и «нарезка» его на неодинаковые фрагменты по 200-500 букв. Задача сборки генома в данном случае сводится к тому, чтобы из получившейся смеси многокопийных участков восстановить исходный текст – геномную последовательность.

Помимо очевидной сложности этой задачи, есть некоторые дополнительные практические проблемы: оригинал может иметь много повторяющихся последовательностей, кроме того, некоторые фрагменты могут быть повреждены во время фрагментации, что способствует возникновению ошибок при чтении. Фрагменты другой чужеродной ДНК

также могут попасть в секвенируемые пробы, что также осложняет процесс восстановления исходной последовательности.

Сложность сборки последовательности определяется двумя основными факторами: количеством прочитанных фрагментов (ридов) и их длиной. Увеличение количества фрагментов позволяет лучше идентифицировать изначальную последовательность за счет увеличения глубины покрытия (количество раз, которое был отсеквенирован каждый нуклеотид), но в то же время, это увеличивает и алгоритмическую сложность. При наличии референсного генома более короткие последовательности можно картировать быстрее. В то же время они усложняют процесс сборки.

В 2004-2005 компанией 454 Life Sciences был доведен до коммерческой жизнеспособности метод пиросеквенирования. Этот новый метод секвенирования генерировал множество коротких чтений (короче, чем при секвенировании по Сэнгеру): около 400-500 п.н.. Его гораздо более высокая производительность и сравнительно низкая стоимость позволили геномным центрам быстрее освоить данную технологию, что, в свою очередь, спровоцировало развитие программ-ассемблеров, которые могли бы эффективно справиться с подобными наборами чтений.

С 2006 года стала доступна технология компании Illumina (ранее Solexa), которая могла генерировать около 100 млн. ридов за один запуск секвенатора. Технология Illumina первоначально была ограничена длиной всего 36 п.н., что делало ее менее пригодной для *de novo* сборки, но вскоре этот недостаток был исправлен, и риды достигли длины свыше 100 п.н.

На сегодняшний день существуют два подхода для сборки геномов – основанный на перекрытии (overlap-layout-consensus), применяемый в основном для длинных фрагментов, а также подоход, использующий алгоритмы, основанные на графах де Брёйна (применяется для коротких фрагментов) [5, 6].

При секвенировании длинных участков ДНК ранее широко применялся метод дробовика. При этом случайным образом производится массированная

выборка клонированных фрагментов ДНК данного организма, на основе которых может быть составлена его геномная библиотека. Эти участки секвенируют обычными методами, например, методом Сэнгера. Затем алгоритмы сборки генома рассматривают полученные фрагменты одновременно, находя их перекрытия (overlap), объединяя их по перекрытиям (layout) и исправляя ошибки в объединённой строке (consensus). Поиск перекрытий обычно ведется путем построения дерева суффиксов и выравнивания фрагментов друг относительно друга. Как правило, это самый медленный этап сборки. Объединение по перекрытиям происходит с помощью графов перекрытия, при этом, из-за ошибок секвенирования или наличия повторов, нередко могут возникать ложные подграфы. В некоторой степени исправление ошибок и корректировка происходят на последнем этапе при совмещении объединенных ридов в контиги. Данные шаги могут повторяться несколько раз в процессе сборки. Такой подход называется Overlap-Layout-Consensus. Он был наиболее распространён до появления методов секвенирования следующего поколения.

С развитием технологий секвенирования следующего поколения получение фрагментов стало на порядок дешевле, но размер их стал меньше (до 150 нуклеотидов), а количество ошибок при чтении увеличилось (до 3%). При сборке таких данных получили распространение методы, основанные на графах де Брёйна [7].

Граф де Брёйна – ориентированный граф с k^n вершинами, соответствующими k^n различных наборов длины n с элементами из $\{0,1,\ldots,k-1\}$, в котором из вершины $(x_1,\ldots,x_n)$ в вершину $(y_1,\ldots,y_n)$ ребро ведёт в том и только том случае, когда $x_i=y_{i-1},(i=2,\ldots,n)$; при этом самому ребру можно сопоставить набор длины

$n+1:(x_1,\ldots,x_n y_n)=(x_1 y_1,\ldots,y_n)$. Для такого графа не проходящие дважды через одно и то же ребро эйлеровы пути соответствуют последовательности де Брёйна с параметрами $n+1$ и k.

Искомая геномная последовательность – кратчайшая строка, которая содержит чтения в качестве подстрок. Она и соответствует эйлерову пути. При этом задача нахождения эйлерова пути в графе де Брёйна решается за время, пропорциональное размеру входных данных [7].

Современные геномные ассемблеры основаны на эвристических или приближенных алгоритмах. Они собирают не целую геномную последовательность, а контиги и скаффолды.

1.2.3 Выравнивание ридов на геном

При сборке генома, если уже имеются известные геномы близкородственных видов, полезным может оказаться метод картирования коротких прочтений (Short-Read Sequence Alignment). Он заключается в определении позиций в референсном геноме, откуда с наибольшей вероятностью могло быть получено каждое конкретное короткое прочтение. В отличие от метода сборки *de novo*, при котором риды собираются вместе для реконструкции до этого неизвестного генома, многие современные проекты имеют «референсный геном» – уже известный геном другого, близкого организма. Также это может быть набор референсных последовательностей.

Существуют два основных подхода к решению задач картирования: с использованием хеш-таблиц и с использованием суффиксных деревьев или массивов. Процесс поиска с помощью хеширования в разы быстрее и является менее затратным, чем классическое выравнивание с помощью динамического программирования. Однако алгоритмы, основанные на хешировании, плохо справляются с повторами. Для решения этого вопроса используются алгоритмы, основанные на суффиксных деревьях и

суффиксных массивах. Преимущество данного подхода, в частности, заключается в том, что повторы не увеличивают время работы алгоритма, так как повторяющиеся участки «схлопываются» в суффиксном дереве.

BLAST (Basic Local Alignment Search Tool) — семейство компьютерных программ, служащих для поиска гомологов белков или нуклеиновых кислот, для которых известна первичная структура (последовательность) или её фрагмент. Используя BLAST, возможно сравнить имеющуюся последовательность с имеющимися в базе данных и найти предполагаемые гомологи. Программа BLAST была разработана в системе Национальных институтов здравоохранения США и опубликована в журнале Journal of Molecular Biology в 1990 [8].

Все выравнивания принято делить на глобальные (последовательности сравниваются полностью) и локальные (сравниваются только определённые участки последовательностей). Программы серии BLAST производят локальные выравнивания, что связано с наличием в различных белках сходных доменов и паттернов. Кроме этого локальное выравнивание позволяет сравнить мРНК с геномной ДНК. В случае глобального выравнивания обнаруживается меньшее сходство последовательностей, особенно их доменов и паттернов.

После введения изучаемой нуклеотидной или аминокислотной последовательности (запроса) на одну из веб-страниц BLAST, она вместе с другой входной информацией (база данных, размер «слова» и др.) поступает на сервер. BLAST создаёт таблицу всех «слов» (это участок последовательностей, который для белков по умолчанию состоит из трёх аминокислот, а для нуклеиновых кислот из 11 нуклеотидов) и сходных «слов».

Затем в базе данных проводится их поиск. Когда обнаруживается соответствие, то делается попытка продлить размеры «слова» (до 4 и более аминокислот и 12 и более нуклеотидов) сначала без гэпов (пробелов), а затем с их использованием. После максимального продления размеров всех

возможных «слов» изучаемой последовательности, определяются выравнивания с максимальным количеством совпадений для каждой пары запрос-последовательность, и полученная информация фиксируется в структуре SeqAlign. Форматер, расположенный на сервере BLAST, использует информацию из SeqAlign и представляет её различными способами (традиционным, графическим, в виде таблицы) [6].

Для каждой обнаруженной в базе данных программами BLAST последовательности необходимо определить, насколько она сходна с изучаемой последовательностью (запрос) и значимо ли это сходство.

При определении сходства ключевым элементом является матрица замен, так как она определяет показатели сходства для любой возможной пары нуклеотидов или аминокислот. В большинстве программ серии BLAST используется матрица BLOSUM62 (Blocks Substitution matrix 62 % identity, блоковая матрица замен с 62 % идентичности). Исключением являются blastn и megablast (программы, которые выполняют нуклеотид-нуклеотидные сравнения и не используют матрицы аминокислотных замен).

С помощью модифицированных алгоритмов Смита-Уотермана или Селлерса определяются все пары сегментов (продленные «слова»), которые нельзя увеличить, так как это приведёт к уменьшению показателей сходства. Такие пары продленных «слов» называются парами сегментов с максимальным сходством (high-scoring segment pairs, HSP).

Теоретически локальное выравнивание может начинаться с любой пары нуклеотидов или аминокислот выровненных последовательностей. Однако HPS, как правило, не начинаются близко к краю (началу или концу) последовательностей. Для коррекции такого краевого эффекта необходимо вычислять эффективную длину последовательностей. В случае последовательностей длиной более 200 остатков происходит нейтрализация краевого эффекта.

1.3 Аннотация генома

1.3.1 Проверка качества и маскировка повторов

Первым шагом на пути аннотации любого генома является выяснение, готова ли сборка к аннотации. Для описания полноты сборки используется несколько сводных статистических показателей, наиболее важным из которых является N50 –это длина такого контига, что все контиги больше него дают в сумме половину длины всех контигов. К другим полезным статистическим показателям относятся средний размер разрывов в скаффолдах, и среднее количество пробелов в скаффолдах. Хотя и не существует строгих правил, сборка с длиной N50 скаффолдов сопоставимой с размерами генов считается приемлемой для аннотации, так как если N50 скаффолдов примерно равна средней длине гена, то около 50% генов будут содержаться в одном скаффолде. Эти полные гены, вместе с фрагментами из остальной части генома, обеспечат значительный ресурс для последующего анализа [8].

Как правило, первым шагом вычислительного этапа аннотации генома является идентификация и маскирование повторов. В геномах эукариот может содержаться очень много повторов, например, геном человека, как полагают, на 47% состоит из повторов [9], и этот процент, вероятно, является нижним пределом. Кроме того, границы этих повторов, как правило, плохо определены; повторы часто являются вставками для других повторов, и, в большинстве случаев, можно обнаружить лишь фрагменты повторов внутри фрагментов других повторов – полные элементы обнаруживаются крайне редко. Повторы усложняют аннотирование генома. Они должны быть определены и аннотированы, но инструменты, используемые для идентификации повторов, значительно отличаются от тех, которые используются для идентификации генов в геноме хозяина.

Определение повторов осложняется тем фактом, что они плохо сохраняются; таким образом, для точного обнаружения повтора обычно необходимо создавать библиотеку повторов исследуемого генома. Доступные

для этого инструменты можно разделить на два класса: инструменты на основе гомологии и инструменты для создания библиотек повторов *de novo*. Однако следует отметить, что инструменты для *de novo* поиска повторяющихся последовательностей могут находить не только мобильные элементы, но и весьма консервативные белок-кодирующие гены, такие как гистоны и тубулины.

После того, как библиотека повторов создана, она может быть использована в сочетании с таким инструментом, как RepeatMasker [10], в котором используется BLAST [11, 12] и Crossmatch для идентификации участков последовательности целевого генома, гомологичных известным повторам. Термин «маскирование» означает замену каждого нуклеотида, определенного как повтор, на «N» или, в некоторых случаях, на прописные «a», «c», «g», «t» – последний процесс известен как «мягкое маскирование» [12]. Шаг по маскированию повторов предоставляет сведения для инструментов выравнивания и предсказания генов на следующих этапах аннотирования. Неудачное маскирование повторов может привести к катастрофическим последствиям для аннотирования генома. Незамаскированные повторы могут породить миллионы ложных выравниваний BLAST, генерируя ложные сведения для аннотации генов. Что еще хуже, многие открытые рамки считывания (ORFs) транспозонов выглядят как настоящие гены хозяина для программ по предсказанию генов, в результате чего, часть ORFs транспозонов, будут предсказаны как дополнительные экзоны, полностью искажая окончательные аннотации генов. Таким образом, качественное маскирование повторов имеет решающее значение для точной аннотации белок-кодирующих генов.

1.3.2 Предсказание кодирующих областей и функциональная аннотация

Когда появились первые программы-предсказатели генов [13] в 1990-х, они произвели революцию в анализе геномов, так как позволили

сравнительно быстро и легко выявлять гены в собранных последовательностях ДНК. Эти инструменты часто называют *ab initio* предсказателями генов, поскольку они используют математические модели, а не внешние подтверждения (такие как белковые выравнивания, например) для идентификации генов и определения их интрон-экзонной структуры.

Большинство программ по предсказанию генов для моделирования нуклеотидной последовательности (генома) используют скрытые марковские модели [14]. Нуклеотидная последовательность представляется как реализация марковского процесса со скрытыми состояниями, генерирующими фрагменты последовательности ДНК определенной длины.

Большим преимуществом *ab initio* программ для поиска генов для аннотации является то, что принципиально они не нуждаются во внешних свидетельствах для определения гена или для определения его интрон-экзонной структуры. Тем не менее, эти инструменты имеют практические ограничения с точки зрения аннотации. Так, большинство программ по поиску генов находят только последовательности, похожие на кодирующие (CDS) и не сообщают о нетранслируемых областях (UTRs) или об альтернативном сплайсинге транскриптов. Обучение также является проблемой – *ab initio* предсказатели генов используют геномные характеристики, специфичные для данного организма, такие как частоты использования кодонов, распределение длин интронов и экзонов, чтобы отличить гены от межгенных регионов и определить интрон-экзонную структуру. Большинство предсказателей генов используют предварительно вычисленные данные, которые содержат такую информацию для нескольких классических геномов, таких как геномы *C. elegans*, *D. melanogaster*, *A. thaliana*, человека и мыши. Если же исследуемый организм не очень тесно связан с модельными организмами, для которых доступны скомпилированные файлы с параметрами, то встает необходимость обучать программу на изучаемый геном, так как даже близкородственные организмы

могут сильно отличаться по распределению длин интронов и экзонов, частотам использования кодонов и GC-контенту.

При достаточном количестве обучающих данных чувствительность на уровне поиска генов для *ab initio* предсказателей может приближаться к 100%. Однако точность предсказанных интрон-экзонных структур, как правило, значительно ниже, около 60-70%. Для получения очень точных предсказаний генов требуется большое количество уже существующих высококачественных моделей генов и практически идеальных геномных сборок; такие наборы данных редко имеются в наличии для немодельных организмов.

Многие из *ab initio* инструментов, таких как TwinScan, FGENESH, Augustus, Gnomon, GAZE и SNAP могут использовать внешние подтверждения для повышения точности своих предсказаний. ESTs, например, могут быть использованы для однозначной идентификации границы экзона. Этот процесс часто относят к предсказанию генов с доказательством (в отличие от *ab initio*). Предсказание генов с доказательством имеет большой потенциал для повышения качества прогнозирования генов в недавно секвенированных геномах, но на практике его достаточно трудно использовать. ESTs и белки должны сначала быть выровнены на геном; данные РНК-секвенирования также должны быть выровнены на геном, если они доступны. Должны быть определены сайты сплайсинга, собранные доказательства должны быть впоследствии обработаны перед тем, как все эти данные будут готовы к использованию программой-предсказателем. На практике это очень большая работа, требующая много специализированного программного обеспечения. По сути, это является одним из главных препятствий, которое пытаются преодолеть пайплайны по аннотации геномов.

Конечной целью усилий по аннотации является синтез выравниваний на основе доказательств с *ab initio* предсказаниями для получения окончательного набора аннотаций генов. Существует почти столько же

стратегий для создания автоматизированных аннотации, сколько и пайплайнов для аннотации, но общим для всех является использование доказательств и свидетельств для того, чтобы повысить точность моделей генов, как правило, через определенное сочетание пред- и пост-обработки предсказаний генов.

1.4 Однонуклеотидные полиморфизмы

Однонуклеотидный полиморфизм, также известный как простой нуклеотидный полиморфизм (SNP, снипы) является вариацией последовательности ДНК, в которой один нуклеотид – T, C или G – в геноме отличается между членами одного вида, либо между гомологичными участками парных хромосом. Например, два секвенированных фрагмента ДНК от различных индивидуумов – AAGCCTA и AAGCTTA содержат разницу в один нуклеотид. В этом случае говорят о наличии двух аллелей. Почти все общие SNP имеют только два аллеля. Геномное распределение снипов не является однородным; SNP чаще встречаются в некодирующих регионах, чем в кодирующих. То есть, в общем случае, там, где естественный отбор «фиксирует» аллель, исключая другие варианты (что является благоприятной генетической адаптацией), снипы появляются реже [15]. Другие факторы, такие как генетическая рекомбинация и скорость мутации, также могут определять плотность распределения SNP в геноме [16].

Плотность SNP может быть предсказана по наличию микросателлитных последовательностей. В частности, участки с длинными $(AT)_n$ повторами являются мощными предикторами плотности снипов, и с большой вероятностью могут быть найдены в регионах с пониженной плотностью SNP и низким GC-составом [17].

Однонуклеотидные полиморфизмы могут встречаться в кодирующих или в межгенных участках генома. Снипы, находящиеся в пределах кодирующей последовательности, не обязательно меняют аминокислотную

последовательность продуцируемого белка, что объясняется вырожденностью генетического кода. В то же время снипы, которые находятся в не кодирующих белок областях, вероятно, могут влиять на сплайсинг генов, связывание транскрипционных факторов и деградацию РНК.

Однонуклеотидные полиморфизмы, наряду с полиморфизмами длин рестрикционных фрагментов, широко используют в качестве молекулярно-генетических маркеров для построения кладограмм молекулярно-генетической систематики на основе дивергенции гомологичных участков ДНК в филогенезе. Наиболее часто в данной области используются спейсеры генов рибосомальной РНК. Ввиду того, что мутации в данных спейсерах не сказываются на структуре конечных продуктов гена (теоретически они не влияют на жизнеспособность), в первом приближении постулируется прямая зависимость между степенью полиморфизма и филогенетическим расстоянием между организмами.

1.5 Кластеризация

Общий вопрос, задаваемый исследователями во многих областях, состоит в том, как организовать наблюдаемые данные в наглядные структуры. Фактически, кластерный анализ является не столько обычным статистическим методом, сколько набором различных алгоритмов распределения объектов по кластерам.

Цель кластерного анализа заключается в поиске существующих структур. В то же время его действие состоит в привнесении структуры в анализируемые данные, то есть методы кластеризации необходимы для обнаружения структуры в данных, которую нелегко найти при визуальном обследовании. На деле эта привнесенная структура может не совпадать с искомой, «реальной». Кластерный метод всегда размещает объекты по группам, которые могут радикально различаться по составу, если

применяются различные методы кластеризации. Ключом к использованию кластерного анализа является умение отличать «реальные» группировки от навязанных методом кластеризации данных [18].

Кластерный анализ позволяет рассматривать достаточно большой объём информации и сильно сжимать большие массивы, делать их компактными и наглядными. К анализируемым данным часто предъявляются два фундаментальных требования — однородность и полнота. Однородность требует, чтобы все кластеризуемые сущности были одной природы, и описываться они должны сходным набором характеристик [19].

Как и любой другой метод, кластерный анализ имеет определенные недостатки и ограничения. В частности, состав и количество кластеров зависит от выбираемых критериев разбиения. При сведении исходного массива данных к более компактному виду могут возникать определённые искажения, а также могут теряться индивидуальные черты отдельных объектов за счёт замены их характеристиками обобщённых значений параметров кластера.

Кластерный анализ может применяться для решения различных типов задач. Например, для цели понимания имеющихся данных, путем выявления в них структуры. Разбиение выборки на группы схожих объектов позволяет упростить дальнейшую обработку данных и принятия решений, применяя к каждому кластеру свой метод анализа. В таком случае число кластеров стараются сделать поменьше. Если же исходная выборка избыточно большая, то можно сократить её, оставив минимальное количество наиболее типичных представителей от каждого кластера. В этом случае важнее обеспечить высокую степень сходства объектов внутри каждого кластера, тогда как их количество роли не играет. Так же, кластеризация применима для выделения нетипичных объектов. Наибольший интерес при этом представляют отдельные объекты, не вписывающиеся ни в один из кластеров.

Для сравнения объектов необходимо иметь критерий, которым, как правило, является расстояние между объектами. Наиболее общим и самым

распространенным является евклидово расстояние, являющееся геометрическим расстоянием в многомерном пространстве данных, вычисляемое по теореме Пифагора. В качестве других метрик могут быть выбраны квадрат евклидова расстояния, манхэттенское расстояние (сумма модулей разностей координат между точками), расстояние Чебышева (максимум модуля разности компонент между векторами.) и т.д.

1.5.1 Метод динамических ядер (*k*-means)

В общем случае среди множества методов кластерного анализа можно выделить два наиболее распространенных – это иерархические и итерационные. Иерархическая кластеризация заключается в упорядочивании данных на основании сходства между объектами, обычно при помощи графов без циклов, построенных по матрице мер близости. Однако, при большом количестве наблюдений иерархические методы не пригодны. В таких случаях используют неиерархические методы, основанные на разделении, которые представляют собой итеративное дробление исходной совокупности. В процессе деления новые кластеры формируются до тех пор, пока не будет выполнено правило остановки.

Одним из наиболее популярных методов кластеризации, наряду с методом главных компонент, обеспечивающим наилучшую аппроксимацию данных, является метод динамических ядер или *k*-means.

Действие алгоритма таково, что он стремится минимизировать среднеквадратичное отклонение точек кластеров от центров этих кластеров.

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2 , \quad (1)$$

где k — число кластеров, S_i — полученные кластеры, $i=1,2,\ldots,k$, μ_i — центры масс векторов $x_j \in S_i$.

Набор k точек $Y=\{y_1,\ldots,y_k\}, y_i \in R^m$, называется набором главных точек для массива данных X, если он аппроксимирует X с минимальной среднеквадратичной ошибкой расстояния по всему набору k-точек в R^m:

$$\sum_{x \in X} dist^2(x, P(x,Y)) \to min, \quad (2)$$

где $P(x,Y)$ это ближайшая к **x** точка из Y.

Сначала выбираются случайные точки: $y_1,\ldots,y_k$ из $x_i \in X$ (с равными вероятностями). Множество элементов X разбивается на подмножества точек данных $K_i, i=1..k$ по их близости к y_k:

$$K_i=\left\{x : y_i=\arg\min_{y_j \in Y} dist(x, y_j)\right\} \quad (3)$$

Алгоритм начинает с k случайно выбранных кластеров, а затем изменяет принадлежность объектов к ним, чтобы минимизировать изменчивость внутри кластеров, и максимизировать изменчивость между кластерами. На каждой итерации перевычисляются центры масс кластеров, полученных на предыдущем шаге:

$$¿\, K_i \vee ¿ \sum_{x \in K_i} x, i=1..k$$
$$y_i=\frac{1}{¿} \quad (4)$$

Затем векторы разбиваются на кластеры вновь в соответствии с тем, какой из новых центров оказался ближе по выбранной метрике [20]. Алгоритм завершается, когда на какой-то итерации не происходит изменения центра масс кластеров. Это происходит за конечное число итераций, так как количество возможных разбиений конечного множества конечно, а на каждом

шаге суммарное квадратичное отклонение не увеличивается, поэтому зацикливание невозможно.

Естественно, результат разбиения сильно зависит как от числа k заданных кластеров, так и от выбора исходных центров кластеров, чей оптимальный выбор неизвестен. Кроме того, алгоритм слишком чувствителен к выбросам, которые могут искажать среднее.

1.5.2 Упругие карты

Метод упругих карт схож с самоорганизующимися картами. Для аппроксимации облака данных используется упорядоченная система вершин, которая размещена в многомерном пространстве.

Определим упругую карту как простой ненаправленный граф $G(Y,E)$ с набором вершин $Y=\{y^{(i)}, i=1..p\}$ и граней $E=\{E^{(i)}, i=1..s\}$. Объединим некоторые смежные грани в пары $R^{(i)}=\{E^{(i)},E^{(k)}\}$ и обозначим как набор элементарных ребер $R=\{R^{(i)}, i=1..r\}$.

Каждая грань $E^{(i)}$ имеет начальную $E^{(i)}(0)$ и конечную $E^{(i)}(1)$ вершину. Элементарное ребро представляет собой пару смежных граней и имеет начальную вершину $R^{(i)}(1)$, конечную вершину $R^{(i)}(2)$, и центральную $R^{(i)}(0)$.

Разместим вершины сети в многомерном пространстве данных. Сделать это можно располагая их случайным образом или разместив вершины в выбранном подпространстве. Например, можно разместить их в подпространстве, «натянутом» на первые две-три главные компоненты. В любом случае каждая вершина графа становится вектором в R^M.

Затем для графа G определим энергию U, которая суммирует энергии каждой вершины, грани и ребра:

$$U = U^{(Y)} + U^{(E)} + U^{(R)} . \quad (5)$$

Разделим всю совокупность точек данных на кластеры (таксоны) $K^{(i)}, i = 1 \dots p.$ Каждый из них содержит точки данных, для которых вершина $y^{(i)}$ является наиболее близкой:

$$K_i = \left\{ x^{(j)} : \left\| x^{(j)} - y^{(i)} \right\| \vec{Y} min \right\} \quad (6)$$

Упругая энергия вложенного графа определяется следующим образом:

$$U^{(Y)} = \frac{1}{N} \sum_{i=1}^{p} \sum_{x^{(j)} \in K^{(i)}} \left\| x^{(j)} - y^{(i)} \right\|^2 , \quad (7)$$

$$U^{(E)} = \sum_{i=1}^{s} \lambda_i \left\| E^{(i)}(1) - E^{(i)}(0) \right\|^2 , \quad (8)$$

$$U^{(R)} = \sum_{i=1}^{r} \mu_i \left\| R^{(i)}(1) + R^{(i)}(2) - 2R^{(i)}(0) \right\|^2 . \quad (9)$$

Здесь $U^{(Y)}$ это средний квадрат расстояния между вершиной $y^{(i)}$ и точками данных в $K^{(i)}$ кластере, $U^{(E)}$ – аналог суммарной энергии упругого растяжения, а $U^{(R)}$ – аналог суммарной энергии упругой деформации сети. Можно представить, что каждая вершина соединена упругими связями с ближайшими точками данных и одновременно со смежными вершинами.

Значения λ_i и μ_i являются коэффициентами растяжимой упругости каждой грани $E^{(i)}$ и коэффициентом упругости изгиба каждого ребра $R^{(i)}$ соответственно.

Упрощенное рассмотрение показывает, что если необходимо, чтобы упругая энергия сетки оставалась неизменной в случае тонкой сетки, то

$$\lambda=\lambda_0 s^{\frac{2-d}{d}}, \quad \mu=\mu_0 r^{\frac{4-d}{d}}, \tag{10}$$

где d – размерность сети ($d=1$ в случае полилинии, $d=2$ в случае шестигранной, прямоугольной и сферической решетки, $d=3$ в случае кубической и т.д.).

Энергия U минимизируется до получения оптимальной конфигурации узлов. Затем сеть используется как нелинейный экран для визуализации распределения точек данных путем проецирования их на многообразия, построенные с использованием самой сети в качестве точки приближения. Далее при помощи окрашивания можно визуально отображать различные функции координат пространства данных [21].

Целью применения эластичной сетки является введение точки аппроксимации в многообразия. Чтобы избежать путаницы, следует заметить, что термин «упругая сетка» был независимо друг от друга введен несколькими группами для разных целей: для решения задачи коммивояжера (Дурбин & Willshaw, 1987), в контексте главных многообразий (Горбань и др, 2001), в контексте регуляризованной задачи регрессии (Zhou & Гесте, 2005). Эти три понятия полностью независимы и имеют разное значение [20].

1.5.3 Частотные словари

Понятие частотного словаря чаще всего используется в лингвистике. Если говорить о нуклеотидных последовательностях, то в качестве алфавита здесь будут выступать нуклеотиды (аденин, тимин, гуанин, цитозин), в качестве текста – геномные последовательности. Тогда под частотным словарем будет пониматься множество всех символьных подпоследовательностей заданной длины, встречающихся в изучаемой последовательности, вместе с указанием частоты их встречаемости [22].

Пусть имеется символьная последовательность из четырёхбуквенного алфавита $\aleph=\{A,C,G,T\}$. Триплетом в данном случае будут назваться три подряд стоящих символа $\nu_1\nu_2\nu_3$. Частотным словарём W_3 будет список всех триплетов (не более 64) с указанием их частот. Все частоты связаны соотношением:

$$\sum_{\nu_1\nu_2\nu_3} f_{\nu_1\nu_2\nu_3}=1 \qquad (11)$$

Частота определяется как отношение числа копий данного триплета, обнаруженных в последовательности, к их общему числу (равному длине всей последовательности). Тем самым каждый геном отображается точкой в 63-мерное пространство частот [23].

При составлении частотного словаря в качестве текста можно брать не только большие массивы из множества геномов, но и фрагменты отдельных геномов организмов. Если принять за «слова» триплеты и подсчитывать частоту их встречаемости во фрагментах, определенных среди одного конкретного генома, то применяя к полученным данным методы кластерного анализа, можно, например, судить о распределении генных участков в данном геноме.

2 Материалы и методы

2.1 Сборка хлоропластного генома лиственницы сибирской

Для сборки генома были использованы данные полногеномного секвенирования *L. sibirica*, полученные на приборе Illumina HiSeq2000 в Лаборатории лесной геномики СФУ под руководством проф. К. В. Крутовского. Образцы ДНК были взяты для трех деревьев: красноярского (Института леса им. В. Н. Сукачева), хакасского (окр. с. Черное озеро, Республика Хакасия) и уральского. Для выделения ДНК у лиственницы сибирской использовали хвою, гаплоидный каллус. В качестве референсных последовательностей из NCBI GenBank были взяты данные для хлоропластных геномов *Larix decidua* Mill. [24] и *Larix occidentalis* Nutt. [25] (AB501189.1 и FJ899578.1). Размер библиотек составил 400-500 п.н. (уральское дерево) и 5000 п.н. (хакасское дерево).

Выравнивание ридов проводилось с использованием программы для картирования коротких сиквенсов на геномы средних размеров Bowtie2. Данная программа базируется на алгоритме построения FM-индекса, основанном на преобразовании Барроуза-Уилера. Ассемблирование производилось при помощи геномного ассемблера SPAdes. Для аннотирования последовательности был использован сервис Rapid Annotation using Subsystem Technology (RAST). Поиск снипов производился с помощью программного обеспечения Ugene (опция Call Variants with SAMtools).

На первом этапе сборки при помощи Bowtie2 производилось картирование имеющихся ридов уральского дерева на известные референсы *L. decidua* и *L. occidentalis* (multi-fasta файл).

```
ebondar@genom-gpu:~/Assembly> bowtie2 --very-sensitive
--threads 10 --no-unal --al-conc Cart_reads.fastq -x
DeciOcci_Hak -1 Larix_h -2 Larix_h2 -S DeciOcci_Hak.sam
```

Картировавшиеся риды затем собирались программой SPAdes. Получившиеся контиги вновь выравнивались с использованием BLASTN на референс *L. decidua.* Отобранные таким образом контиги получили статус «правильных». Далее вновь проводилось ассемблирование с помощью SPAdes: «правильные контиги» подавались на вход с указанием флага «--trusted contigs».

```
spades.py -1 Cart_reads.1.fastq -2 Cart_reads.2.fastq
--trusted-contigs contigs.fasta -o Sp2_reads -t 20
```

Заключительный этап сборки – формирование скаффолдов осуществлялось программой SSPACE. Для этого использовались полученные контиги и библиотеки ридов по лиственнице хакасского дерева. В результате был получен один скаффолд длиной 122561 п.н.

```
perl SSPACE_Standard_v3.0.pl -l libraries.txt -s
contigs.fasta
```

2.2 Поиск кодирующих участков и аннотирование

Как известно, молекула хлоропластной ДНК имеет кольцевую структуру. Собранный же скаффолд является линейной последовательностью. Чтобы выяснить местонахождение «первой позиции» в собранном геноме и соответственно ориентировать последовательность, было произведено картирование сборки на референс *L. occidentalis*. Оказалось, что первому нуклеотиду в референсе соответствует 87189 позиция в собранном геноме. Из чего можно сделать вывод, что сборка «сдвинута» относительно референса на 87189 п.н. На основании этих данных, от начала файла был отрезан фрагмент длиной в 87189 п.н. и вставлен в конец последовательности. Проверка собранного генома при помощи RepeatMasker повторов не выявила.

Для определения местонахождения кодирующих участков в собранном хлоропластном геноме первоначально был использован

GeneMarkS – предиктор, с самообучающимся алгоритмом, основанный на применении скрытых марковских моделей. После запуска на наших данных предиктор выявил 72 кодирующих участка, что значительно меньше, чем у сходного по длине хлоропластного генома *L. decidua*, в котором найдено 110 кодирующих зон [24].

Предполагается, согласно теории симбиогенеза, что хлоропласты произошли от цианобактерий около 1–1,5 млрд лет назад [26]. Следовательно, геномы хлоропластов очень близки к бактериальным геномам. В связи с этим для аннотирования генома был выбран сервис Rapid Annotation using Subsystem Technology (RAST) [27], предназначенный для аннотирования бактериальных и архейных геномов.

Полученная аннотация содержала как подтвержденные участки, с известными генами, так и участки, которые предиктор распознал как потенциально кодирующие (hypothetical protein). Для того чтобы уточнить функции этих гипотетических кодирующих участков проводилось сравнение с аннотациями близкородственных видов *L. decidua* и *L. occidentalis*., а так же выборочное картирование фрагментов генома при помощи BLASTN. Участки hypothetical protein, подтвержденные BLASTN как кодирующие, именовались соответствующим образом. Всего был выявлен 121 кодирующий участок.

2.3 Поиск однонуклеотидных полиморфизмов

Для поиска однонуклеотидных полиморфизмов использовались программы Bowtie2 и Ugene (Call Variants with SAMtools) [28]. Поиск производился между тремя деревьями: уральским, хакасским (Черное озеро), и красноярским (Институт леса им. Сукачева). Риды красноярского и хакасского деревьев выравнивались на окончательную сборку уральского дерева.

Результаты выравнивания – в полученном sam-файле – вместе с нуклеотидной последовательностью для уральского дерева использовались

программой Ugene для поиска снипов. В итоге для трех деревьев было найдено 8 позиций с SNP.

2.4 Построение частотных словарей

Полученная в ходе ассемблирования нуклеотидная последовательность хлоропластного генома использовалась для составления частотных словарей.

Выделение фрагментов и подсчет частот осуществлялись при помощи скрипта, написанного на языке Perl. Составление частотного словаря хлоропластного генома *L. sibirica* происходило следующим образом: в исходном геноме последовательно выделялись фрагменты заданной длины $w = 303$ п.н.. Каждому такому фрагменту присваивался номер $S_i = i \cdot p + w/2 + 0.5$, который обозначает позицию центрального нуклеотида i-го фрагмента в геноме, начиная с его начала. Фрагменты выделялись с шагом $p = 10$ п.н., равным числу нуклеотидов между двумя соседними S_i позициями (центрами) фрагментов.

После этого для каждого фрагмента составлялся частотный словарь: были взяты триплеты, расположенные последовательно друг за другом и не пересекающиеся между собой; для каждого триплета была посчитана частота его встречаемости в каждом отдельно взятом фрагменте. При этом было произведено нормирование данных: полученные абсолютные частоты были превращены в относительные путем деления их на сумму частот всех триплетов фрагмента. Получаемый текстовый файл с результатами в виде

перечисления фрагментов, их позиций S_i, а также 64-х триплетов и их частот, конвертировался в формат DAT.

В любой последовательности существует всего 3 положения рамки считывания: нулевой сдвиг (т.е. старт от начала фрагмента), сдвиг на один нуклеотид и сдвиг на два. Так как наш исходный геном делится на фрагменты, отстоящие друг от друга с шагом $p = 10$, то каждый следующий фрагмент будет сдвинут по сравнению с предыдущим на один нуклеотид относительно положения триплетов. Таким образом, все вырезанные фрагменты последовательно будут иметь фазу 0, 1, или 2, если ее считать от начала последовательности (абсолютная фаза в геноме). Для удобства кластеризации в исходный DAT-файл так же были добавлены данные о фазе каждого фрагмента.

2.5 Классификация методом динамических ядер

При разбиении один из шестидесяти четырёх триплетов исключался из рассмотрения, чтобы линейная связь не создавала ложного сигнала.

В массиве данных (12226 фрагментов) определялся порядок. Для этого строилась классификация методом динамических ядер (*K*-means) с разбиением на 2, 3, 4, 5, 6 и 7 классов по 50 итераций каждое; всюду использовалось Евклидово расстояние. Для всех фрагментов подсчитывалось количество раз, когда они переходили из класса в класс. Если суммарное число переходов из одного класса в другой было не более 5, то таким фрагментам присваивался статус «устойчивых». Остальные – те, что часто мигрировали из класса в класс – считались неустойчивыми.

2.6 Абсолютная и относительная фаза фрагмента в геноме

Разделение генома на фрагменты имеет строго заданный порядок (длина 303 с шагом в 10 и, соответственно, перекрытием в 293 нуклеотида).

Абсолютная фаза в геноме была задана искусственно: от начала генома до его конца каждый сдвиг гипотетической рамки считывания принимался за значения 1, 2 или 3, следующие в повторяющемся порядке. При помощи визуализации в главных компонентах было произведено разделение фрагментов в зависимости от того, к какой абсолютной фазе они принадлежат.

Для каждого из фрагментов (для которых строились частотные словари) введем понятие относительной фазы. Относительная фаза будет определять положение фрагмента не относительно начала всей последовательности (абсолютная фаза), а относительно того кодирующего участка (гена), в который этот фрагмент попал. Дело в том, что ген может располагаться от стартовой позиции генома на расстоянии не кратном трем. То есть, фрагменты, попавшие в один и тот же кластер, могут иметь разные абсолютные фазы, но при этом их относительные фазы, скорее всего, окажутся одинаковы в пределах своего кластера.

В каждом из кластеров была выделена группа фрагментов (25-35 штук), составляющих его центр. После этого каждый такой фрагмент (точнее его S_i-идентификатор описывающий положение фрагмента в геноме) сопоставлялся с геномной аннотацией для определения его положения — находится ли он в кодирующей области или в межгенном промежутке. Из координаты S_i-фрагмента вычиталась координата стартовой позиции гена, и от полученной новой S_i-координаты (относительно стартовой позиции гена) находился остаток от деления на 3. Таким образом, относительная фаза фрагмента – это остаток от деления на три порядкового номера нуклеотида, соответствующего центру фрагмента, отсчитанного от начала гена, в котором этот фрагмент находится.

2.7 Построение упругой карты

Кластеризация методом упругой карты была проведена при помощи программы VidaExpert [29].

Каждый фрагмент генома имеет набор характеристик – частоты встречаемости шестидесяти четырёх триплетов, подсчитанные в нем. От этих частот зависит распределение фрагментов относительно друг друга в 64-мерном пространстве. Однако при построении эластичной карты один из шестидесяти четырёх триплетов исключался из рассмотрения. Делалось это потому, что сумма частот всех 64 триплетов всегда равна 1. Такая линейная связь, если её не исключить, создаёт ложный сигнал. Формально исключить можно любой триплет, однако на практике исключался тот, для которого стандартное отклонение по всем фрагментам генома было минимальным. Для данного генома таким является триплет CGC. Таким образом, каждый фрагмент генома при визуализации отображается точкой в 63-мерном пространстве частот.

Поскольку геном проаннотирован — то есть известно расположение кодирующих и некодирующих участков в нем — не трудно провести сравнение и узнать, к каким областям относятся фрагменты того или иного кластера.

Также, из исходного DAT-файла, содержащего информацию о частотных словарях S_i -фрагментов, были исключены все фрагменты, S_i -идентификаторы которых, в соответствии с аннотацией, принадлежали межгенным промежуткам. На основании этих данных была вновь проведена кластеризация методом упругой карты.

Для кластеров, по предварительной оценке отнесенных к «генным», был определен их более точный генный состав (конкретные гены, которым принадлежат попавшие в них фрагменты). S_i -идентификаторы фрагментов из данных трех кластеров сравнивались с аннотацией генома – таким образом определялось, к какому гену относится каждый фрагмент.

3 Результаты

3.1 Сборка, аннотирование и поиск SNP

Длина хлоропластного генома *L. sibirica* составила 122561 п.н. и близка к 122474 п.н. у близкородственной *Larix decidua* Mill., 1768.

В результате аннотирования и сравнения полученных данных с уже имеющимися близкородственными видами *L. decidua* и *L. occidentalis*, был выявлен 121 кодирующий участок. Из них 87 соответствуют белок-кодирующим генам и 34 – генам (Таблица 1 в Приложении). Карта расположения генов в геноме показана на Рисунке 1, характеристика генома в Таблице А4 в Приложении.

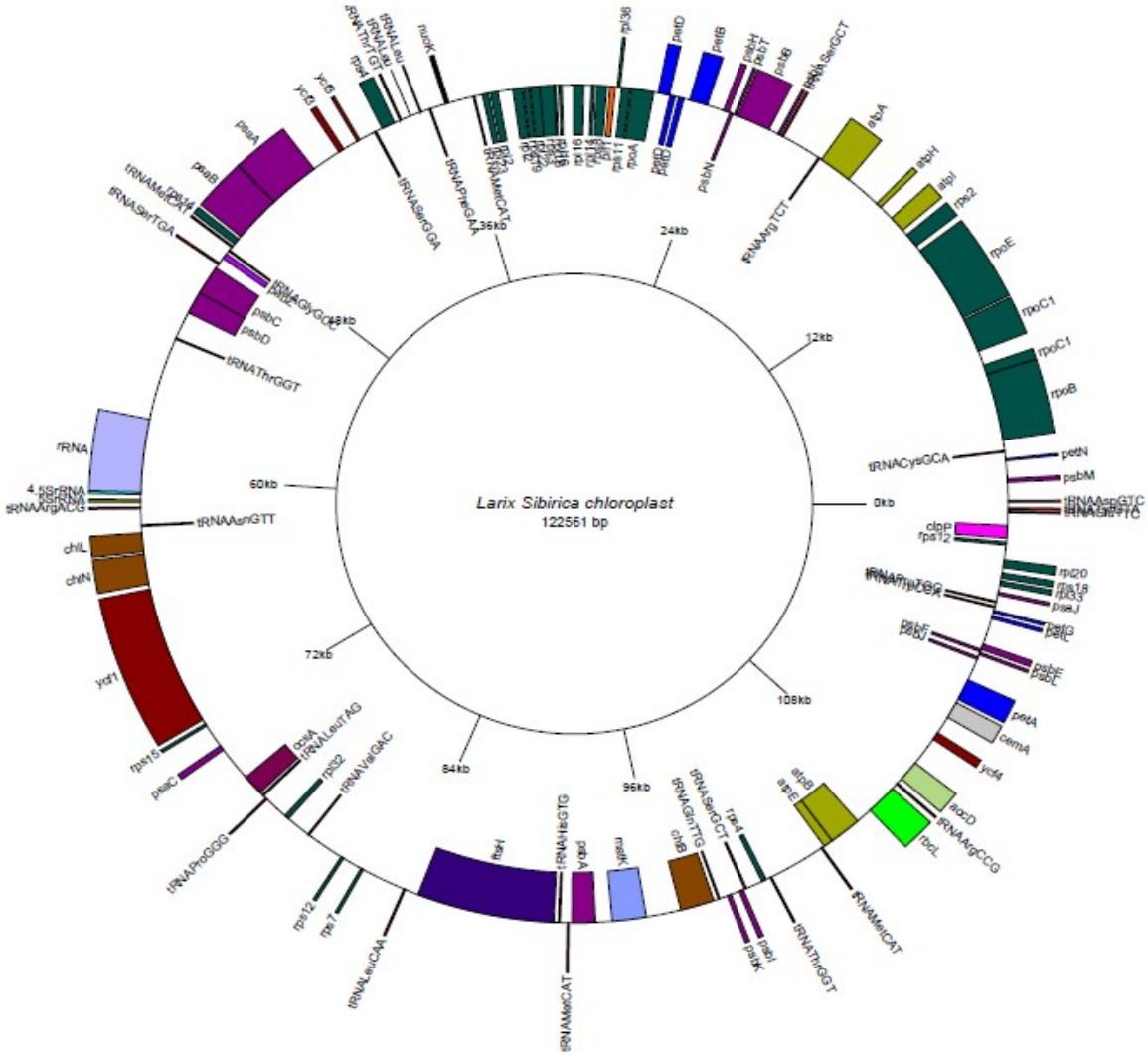

Рисунок 1 – Карта расположения генов в кольцевом геноме хлоропласта *L. sibirica*.

Для трех деревьев найдено 8 позиций с SNP (Таблица 2), 2 из них находятся в кодирующих участках генома*: hypothetical protein* и *Cell division protein FtsH.*

Таблица 2 – Однонуклеотидные полиморфизмы, найденные для трех деревьев, произрастающих в Красноярске, Хакасии и на Урале.

Позиция нуклеотида в геноме	Референс (уральское дерево)	Красноярское/хакасское деревья	Ген
19000	T	.	*hypothetical protein*
37101	T	A	-
37102	G	T	-
37103	G	A	-
56509	C	A	-
78980	.	A	-
90976	T	C	*Cell division protein FtsH*
91006	C	.	-

3.2 Кластеризация методом k-means

При классификации методом динамических ядер были получены варианты разбиения на 2, 3, 4, 5, 6 и 7 классов. Результаты приведены в Таблице 3. Визуальное распределение по классам изображено на Рисунке 2. Наиболее устойчивым здесь является разбиение на два класса.

Таблица 3 – Структура устойчивости при разбиении фрагментов методом динамических ядер. K — число классов, на которое проводилось разбиение; U — число неустойчивых фрагментов в разбиении.

K	2	3	4	5	6	7
U	6911	7983	7757	7787	7897	8665

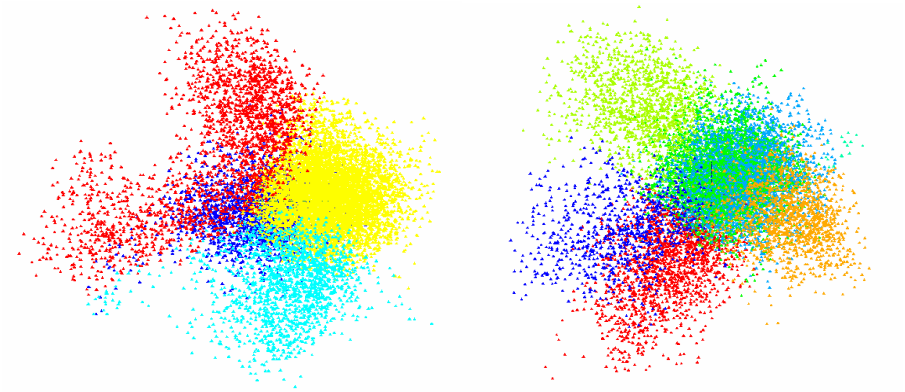

Рисунок 2 – Распределение фрагментов по классам, полученным методом динамических ядер: слева на 4 класса, справа на 7 классов. Изображения приведены в координатах, соответствующим главным компонентам.

Результат разделения фрагментов в зависимости от того, к какой абсолютной фазе они принадлежат, показан на Рисунке 3. Видно, что все фрагменты, имеющие разные абсолютные фазы равномерно распределены по классам, выделяемым методом динамических ядер. Это обусловлено тем, что структурно упорядоченные участки в геноме не следуют столь строгой периодичности, что была искусственно задана нами.

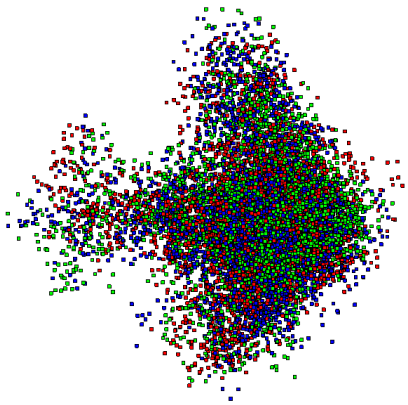

Рисунок 3 – Распределение по фазам фрагментов в хлоропластном геноме лиственницы сибирской. Визуализация методом PCA. Тремя разными цветами выделены фрагменты, попадающие в разные фазы в геноме – 0, 1 и 2.

3.3 Кластеризация методом упругих карт

В результате кластеризации при помощи упругой карты (Рисунок 4) было получено 7 кластеров при исключении триплета GCG (обладающего наименьшим стандартным отклонением) и 8 кластеров при исключении

триплета GCA. Для определения «качественного состава» кластеров были взяты именно два данных варианта разбиения.

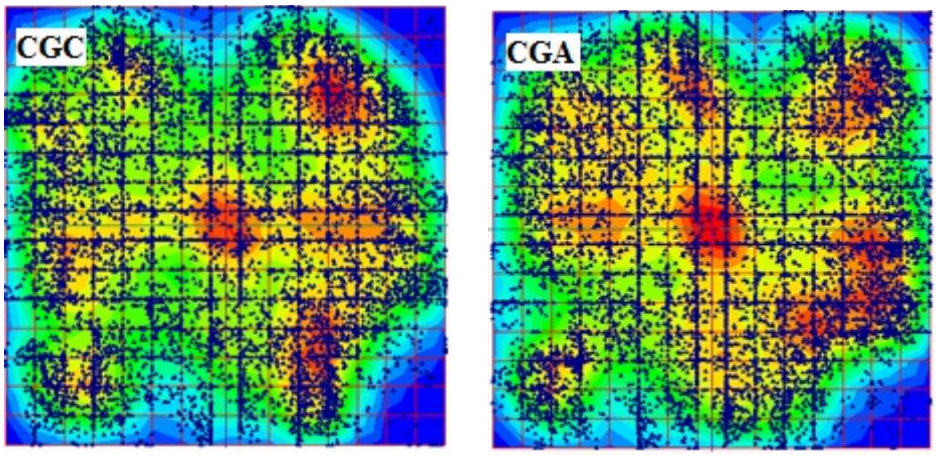

Рисунок 4 – Кластеризация методом упругих карт. Визуализация во внутренних координатах. Слева семь кластеров, полученных при исключении триплета GCG, справа – восемь кластеров триплета CGA.

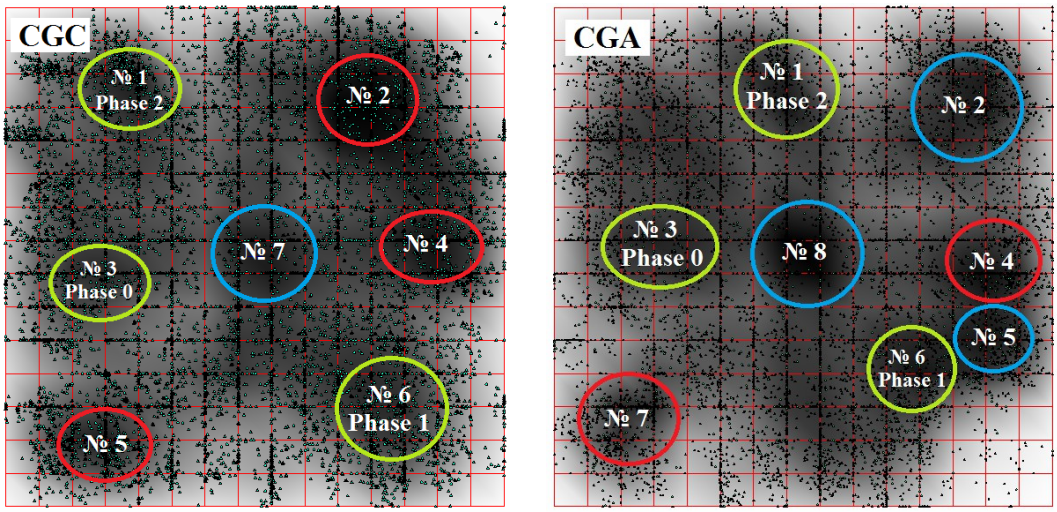

Рисунок 5 – Соотношение генных и межгенных областей в каждом из кластеров. Слева – при исключении триплета CGC, справа – при исключении триплета CGA. Зеленым цветом обозначены кластеры, состоящие из только и почти только «генных» фрагментов, красным цветом – «межгенные» кластеры, синим – кластеры с примерно равным соотношением генных и межгенных фрагментов. Для «генных» кластеров обозначена собственная фаза, к которой относятся все входящие в него фрагменты.

В результате сравнения аннотации и S_i -координат фрагментов были получены «качественные характеристики» кластеров — оценочное соотношение генных и межгенных областей в каждом из них (Рисунок 5).

Центральный кластер получился смешанным для обоих (CGA и CGC) исключенных триплетов. Кластеры № 1, № 3 и № 6 в обоих случаях состоят из только и почти только «генных» фрагментов (зеленым цветом на картинках). Красным цветом обозначены «межгенные» кластеры (65-90% «межгенных» фрагментов), синим — с примерно равным соотношением генных и межгенных фрагментов.

Так же были определены относительные фазы фрагментов — относительно генов, в которых они находятся (для каждого кластера набор генов свой). Относительные фазы фрагментов в рамках своих генов одинаковы, это наблюдается для всех генов во всех кластерах. Кроме того, для кластеров, состоящих только из генных фрагментов характерно «единство фазы»: последние в большинстве своем имеют одинаковые фазы, независимо от того, в каком гене находятся (внутри конкретного кластера). Таким образом, в кластер № 1 вошли фрагменты с относительной фазой 2, в кластер № 3 – с относительной фазой 0, в кластер № 6 – с относительной фазой 1 (Рисунок 6).

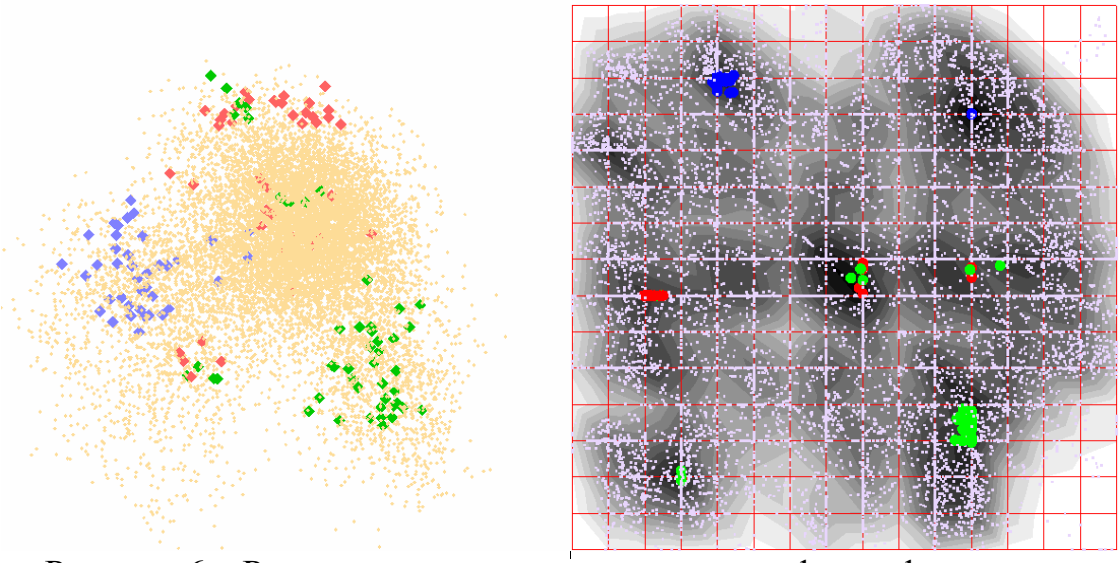

Рисунок 6 – Распределение по относительным фазам фрагментов в хлоропластном геноме лиственницы сибирской. Визуализация методом PCA (слева) и во внутренних координатах (справа). Тремя разными цветами выделены фрагменты, попадающие в разные фазы в геноме – 0 (красный), 1 (зеленый) и 2 (синий).

В результате разбиения фрагментов, входящих лишь в кодирующие области генома, среди генных фрагментов выделилось два кластера (Рисунок 7).

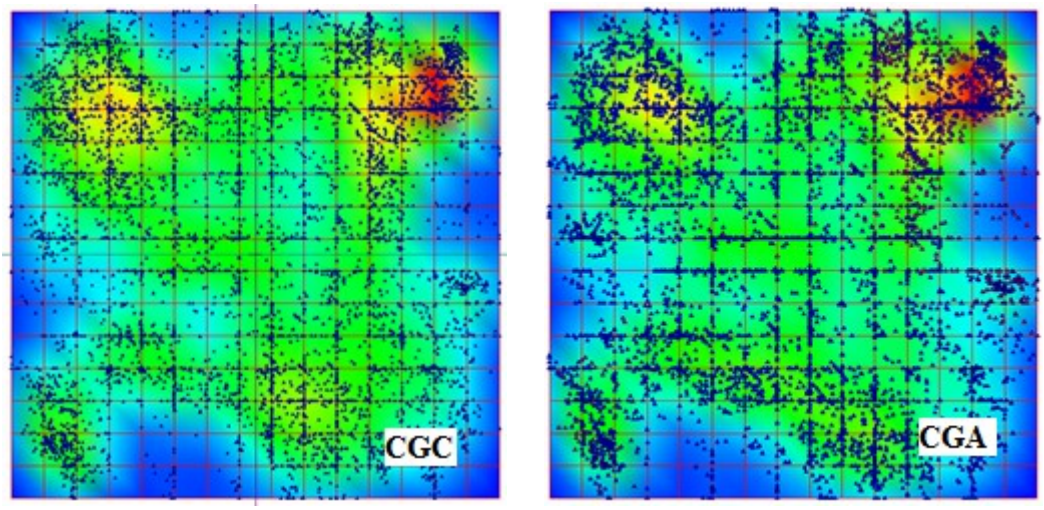

Рисунок 7 – Разделение на кластеры только генных фрагментов методом упругой карты. Визуализация во внутренних координатах. Слева исключен триплет CGC, справа – триплет CGA.

Всего в «генные» кластеры попали 44 гена (Рисунок 8). Для всех трех кластеров общими оказались 16 генов (Таблица А2 в Приложении). Еще 9 генов общие для двух из трех кластеров, остальные 19 – разные для всех трех (Таблица А3 в Приложении).

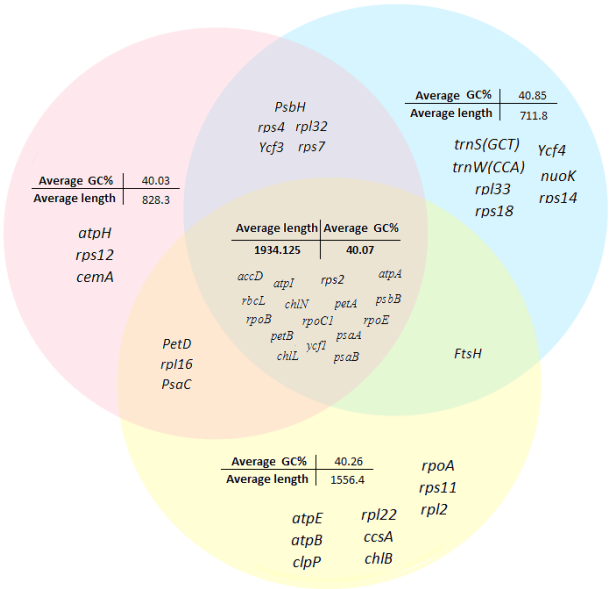

Рисунок 8 – Распределение генов по кластерам. Каждый круг соответствует одному из генных кластеров.

4 Обсуждение результатов

Исследования геномов органелл проводятся во многих биоинформатических лабораториях в мире. Из геномов хвойных растений наиболее изученными можно считать род *Pinus*. Для рода *Larix* опубликовано лишь два хлоропластных генома.

В данной работе впервые была проведена сборка и аннотирование хлоропластного генома лиственницы сибирской. Сравнительный анализ хлоропластных геномов сибирской и европейской лиственниц показал значительное сходство их генной структуры. Кластерный анализ частотного словаря генома так же показал определенное сходство со структурой бактериальных геномов.

Для бактериальных геномов описана семикластерная структура, состоящая из двух треугольников, вершины которых являются кластерами [30]. Взаимное расположение треугольников – и кластеров соответственно – в значительной степени зависит от GC-состава генома. Для геномов с близким к равновесному (0,40 – 0,46) GC-составом вершины треугольников располагаются противоположно относительно друг друга и образуют семикластерную структуру. При смещении GC-состава в ту или иную сторону, происходит вырождение структуры в четырехкластерную (с проекцией вершин друг на друга). Согласно теории симбиогенеза, предками хлоропластов являются цианобактерии, из чего можно предположить, что и структуры их геномов будут подобны или близки. Количество кластеров и рисунок их расположения у хлоропласта имеет сходство с семикластерным расположением у бактерий, хотя GC-состав его генома составляет 38,75%.

В бактериальных геномах два треугольника кластерной структуры содержат фрагменты кодирующих областей генома из прямого и обратного стренда, соответственно. Центральный кластер содержит фрагменты, находящиеся в некодирующих участках генома.

По составу кластеров полученная семикластерная структура хлоропластного генома лиственницы сибирской отличается от ранее описанной структуры бактериальных геномов. В хлоропластном геноме центральный кластер оказался смешанным по составу, тогда как кодирующие и некодирующие области оказались отнесены к вершинам двух треугольников.

Причина такого отличия генома хлоропласта лиственницы от бактериальных геномов не ясна. Различия в GC-составе между кластерами выявлено не было. Возможно, дело в соотношении в геноме генных и некодирующих участков. У бактерий доля некодирующих участков в геноме очень мала. В хлоропластном геноме лиственницы соотношение генных и межгенных участков приблизительно 1:1. Так же свое влияние может оказывать длина генома – хлоропластный геном ощутимо короче среднего бактериального генома.

ПРИЛОЖЕНИЕ

Таблица А.1 – Группы генов хлоропластного генома *L. Sibirica*

Группа генов	Количество кодирующих участков в геноме
tRNA	31
rRNA	3
Ribosomal proteins	25
Photosystems I,II	22
Cytochrome (b6 – f, b559)	10
RNA polymerase	5
ATP synthase	4
Light-independent protochlorophyllide reductase	3
Translation initiation factor 1	1
NAD(P)H-quinone oxidoreductase	1
Cell division protein FtsH	1
Maturase K	1
Ribulose bisphosphate carboxylase	1
Acetyl-coenzyme A carboxyl transferase	1
ATP-dependent Clp protease	1
Hypothetical protein	10

Таблица А.2 – Гены, общие для трех кодирующих кластеров.

Гены	CG%	Длина гена (п.н.)
rpoB	38.43	3137
rpoC1	39.8	2056
rpoE	37.94	3668
rps2	39.04	665
atpI	39.44	752
atpA	41.08	1523
psbB	42.63	1526
petA	40.63	821
psaB	42.44	2339
psaB	41.09	2204
chlL	39.04	875
chlN	39.56	1412
ycf1	36.79	6560
rbcL	44.68	1427
accD	38.05	974
petA	40.48	1007

Таблица А.3 – Гены, разные для двух и трех кластеров.

Кластер № 1 (фаза 2)	Кластер № 3 (фаза 0)	Кластер № 6 (фаза 1)	CG, %
PsbH	-	PsbH	41.23
rps4	-	rps4	37.13
Ycf3	-	Ycf3	40.53
rpl32	-	rpl32	35.68
rps7	-	rps7	42.28
PetD	PetD	-	37.04
rpl16	rpl16	-	42.52
PsaC	PsaC	-	43.09
-	FtsH	FtsH	33.67
atpH	-	-	45.12
-	-	trnS(GCT)	55.68
-	-	nuoK	40.35
-	rpoA	-	38.59
-	rps11	-	43.26
-	rpl22	-	40.48
-	rpl2	-	41.27
-	-	rps14	44.00
-	ccsA	-	38.84
rps12	-	-	40.31
-	chlB	-	38.39
-	atpE	-	41.03
-	atpB	-	42.83
cemA	-	-	35.37
-	-	Ycf4	38.89
-	-	trnW(CCA)	50.00
-	-	rpl33	36.23
-	-	rps18	35.34
-	clpP	-	42.39

Таблица А.4 – Некоторые параметры генома лиственницы сибирской.

Наименование показателя	Значение показателя
Длина генома, п.н.	122561
CG-состав генома, %	38,75
Средняя длина кодирующей области, п.н.	603
Суммарная длина всех кодирующих областей, п.н.	68307
Минимальная длина кодирующей области, п.н.	70
Максимальная длина кодирующей области, п.н.	6560
Количество генов в прямом стренде	69
Количество генов в обратном стренде	52

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Семериков В. Л., Семерикова С. А., Дымшакова О. С., Зацепина К. Г., Тараканов В. В., Тихонова И. В., Экарт А. К., Видякин А. И., Жамьянсурен С., Роговцев Р. В., Кальченко Л. И.. Полиморфизм микросателлитных локусов хлоропластной ДНК сосны обыкновенной (Pinus sylvestris L.) в Азии и восточной Европе // Генетика, 2014. Т. 50. No 6. С. 660-669.
2. Elaine R. Mardis. Next-Generation Sequencing Platforms. Annu. Rev. Anal. Chem. 2013. 6:287-303
3. Ewing B, Hillier L, Wendl MC, Green P (1998). "Base-calling of automated sequencer traces using phred. I. Accuracy assessment". Genome Res. 8 (3): 175–185.
4. Michael A Quail. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 2012, 13:341
5. Zhenyu Li et al. (2012). «Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph». Briefings in Functional Genomics 11 (1): 25-37.
6. Miller JR, Koren S, Sutton G. (2010). «Assembly algorithms for next-generation sequencing data». Genomics 95 (6): 315-327.
7. Pavel A. Pevzner, Haixu Tang, Michael S. Waterman (2001). «An Eulerian path approach to DNA fragment assembly». PNAS 98 (17): 9748-9753.
8. Ye, L. et al. A vertebrate case study of the quality of assemblies derived from next-generation sequences. Genome Biol. 12, R31 (2011).
9. Lander, E. S. et al. Initial sequencing and analysis of the human genome. Nature 409, 860–921 (2001).
10. RepeatMasker 3.0 [Электронный ресурс]: Smit, A. F., Hubley, R. Green, P. – Режим доступа: http://www.repeatmasker.org/webrepeatmaskerhelp.html
11. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. J. Mol. Biol. 215, 403–410 (1990).
12. Korf, I., Yandell, M. & Bedell, J. BLAST: an Essential Guide to the Basic Local Alignment Search Tool 339 (O'Reilly & Associates, 2003).

13. Burge, C., Karlin, S. Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268, 78–94 (1997).
14. Borodovsky M. and Ekisheva S. Problems and Solutions in Biological Sequence Analysis. Cambridge University Press, 2006
15. Barreiro LB; Laval G; Quach H; Patin E; Quintana-Murci L. (2008). "Natural selection has driven population differentiation in modern humans".Nature Genetics 40 (3): 340–345.
16. Nachman, Michael W. (2001). "Single nucleotide polymorphisms and recombination rate in humans". Trends in genetics 17 (9): 481–485
17. M.A. Varela & W. Amos (2010). "Heterogeneous distribution of SNPs in the human genome: Microsatellites as predictors of nucleotide diversity and divergence". Genomics 95 (3): 151–159.
18. Факторный, дискриминантный и кластерный анализ: Пер. с англ. / Дж.-О. Ким [и др.]; под ред. И. С. Енюкова. — Москва : Финансы и статистика, 1989.— 215 с.
19. Жамбю М. Иерархический кластер-анализ и соответствия / М. Жамбю. – М.: Финансы и статистика, 1988. –342 с.
20. Gorban, A. and Zinovyev, A.: Elastic Principal Graphs and Manifolds and their Practical Applications / A. Gorban, A. Zinovyev // Computing – 2005 – №75 – С. 359–379.
21. Gorban, A.N., Zinovyev, A.Yu., and Wunsch, D.C. / Application of the method of elastic maps in analysis of genetic texts. // In: Proceedings of International Joint Conference on Neural Networks (IJCNN Portland, Oregon, July 20-24) (2003)
22. Sadovsky M.G. Information Capacity of Nucleotide Sequences and Its Applications / M.G. Sadovsky // Bulletin of Mathematicla Biology – 2006
23. Садовский М.Г. О фундаментальной связи геномов митохондрий с геномами организмов-носителей / М.Г. Садовский // Биологические науки – 2014 – № 9 – 781-783.
24. C.S.Wu et al. (2011) Comparative chloroplast genomes of pinaceae: insights into the mechanism of diversified genomic organizations, Genome Biol Evol, 3: 309-319.

25. M.Parks et al. (2009) Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes, BMC Biol., 7: 84.
26. Белякова Г. А. Водоросли и грибы // Ботаника: в 4 т. / Белякова Г. А., Дьяков Ю. Т., Тарасов К. Л. — М.: Издательский центр «Академия», 2006. — Т. 1. — 320 с.
27. Ramy K Aziz, Daniela Bartels at al.(2008) The RAST Server: Rapid Annotations using Subsystems Technology, BMC Genomics, 9:75.
28. Okonechnikov K.. *et al* (2012). Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics 28 (8).
29. ViDaExpert v 1.2 [Электронный ресурс] : Institute of Curie // Andrei Zinovyev. — Режим доступа: http://bioinfo-out.curie.fr/projects/vidaexpert/.
30. Gorban A. N. et al. Four basic symmetry types in the universal 7-cluster structure of microbial genomic sequences. // In Silico Biology – 2005 – 5: 25 – 37.