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

КЛАСТЕРНЫЙ АНАЛИЗ РЕЗУЛЬТАТОВ *DE NOVO* СБОРКИ ГЕНОМА ЛИСТВЕННИЦЫ СИБИРСКОЙ

Руководитель: _____ д. ф.-м. н. М. Г. Садовский

Выпускник: _____ С. В. Новикова

Красноярск 2016

Оглавление

ВВЕДЕНИЕ

Существует достаточно большая группа организмов, работа с геномом которых весьма затруднительна. Например, хвойные, существенно отличающиеся от модельных растений. Геном хвойных обладает большой фенотипической пластичностью, богат повторяющимися нуклеотидными последовательностями и, как следствие, имеет большой размер – 12-30 Gb (миллиардов нуклеотидных оснований). По последним данным, геном хвойных может включать в себя до 82 % повторяющихся последовательностей [1].

Очень часто алгоритмы обработки геномных данных не рассчитаны на такой большой объем информации и, как следствие, такие программы требуют огромных вычислительных мощностей, тратят большое количество времени на работу, либо не работают вовсе. В данной работе рассматривалось применение кластеризации как способ предобработки данных, который бы отчасти смог решить эти проблемы.

Кластеризация как метод несет в себе три функции [2]:

1. *Понимание данных путем выявления кластерной структуры.* Для каждого обнаруженного кластера можно применить особый метод анализа.
2. *Сжатие данных.* Дальнейшая работа только с интересующими нас кластерами позволит существенно снизить размерность данных. Также возможна работа только с характерными представителями каждого кластера.
3. *Обнаружение ранее не изученных объектов и закономерностей.* Нетипичные по своей структуре последовательности скорее всего не будут принадлежать ни к одному из кластеров, либо составлять отдельный малый кластер, заслуживающий специального изучения.

Все три пункта позволят упростить работу с геномными данными, особенно если эти данные велики и ранее не аннотированы.

Тема этой работы «Кластерный анализ результатов *de novo* сборки генома лиственницы сибирской».

Целью настоящей работы является поиск структурных групп контигов генома лиственницы сибирской на основе кластеризации.

Задачи проведенного исследования:

1. Выбрать подходящий для наших данных метод классификации и меру расстояния;
2. Освоить метод динамических ядер (k -means), метод упругих карт;
3. Выявить структурно обособленные группы контигов генома лиственницы сибирской;
4. Проанализировать выявленные группы на предмет функциональной обособленности;
5. Визуализировать данные и результаты;
6. Оценить структуру используемой геномной сборки.

ОСНОВНАЯ ЧАСТЬ

1 Обзор литературы

1.1 Методы кластеризации и меры расстояний

Под кластеризацией мы понимаем процесс разбиения множества последовательностей составляющих геном (либо его фрагменты) на классы, при котором элементы, объединяемые в один класс, имеют большее сходство, нежели элементы, принадлежащие разным классам. Под «сходством» в нашем случае подразумевается близость объектов в многомерном пространстве признаков, а задача сводится к поиску скоплений таких объектов.

Синонимами термина «кластеризация» являются «автоматическая классификация» и «обучение без учителя». Кластеризация отличается от классификации тем, что перечень групп четко не задан и определяется в процессе работы алгоритма, а также тем, что деление на группы происходит на основе внутренних свойств данных, а не внешних признаков.

Решение задачи кластеризации принципиально неоднозначно [3]:

- Результат зависит от точности постановки задачи кластеризации;
- существует много эвристических методов кластеризации;
- число кластеров, как правило, неизвестно заранее;
- результат кластеризации существенно зависит от метрики, которая задается субъективно, и от входного набора данных.

В настоящее время разработано более сотни различных алгоритмов кластеризации. Общепринятой классификации методов кластеризации не существует, но ниже представлен ряд групп подходов, объединенных согласно [4].

Вероятностный подход: EM-алгоритм, Дискриминантный анализ, алгоритмы, основанные на непараметрических оценках плотности. Предполагается, что каждый объект общей совокупности данных

принадлежит одному из K классов. Объекты выбираются из генеральной совокупности случайно и независимо друг от друга. Необходимо определить наиболее правдоподобные значения параметров, восстановив закон распределения для каждого класса.

Подход, использующий аналогию с центром тяжести: метод динамических ядер (k -means), k -medians, Алгоритмы семейства FOREL. Для каждой группы определяется вектор средних значений показателей, интерпретируемый как «центр тяжести» группы. Используется критерий внутригруппового рассеяния.

Подходы на основе систем искусственного интеллекта: Метод нечеткой кластеризации C -means, Нейронная сеть Кохонена, Генетический алгоритм. Типичная архитектура представляет собой однослойную сеть, в которой каждый нейрон соответствует некоторому кластеру. В процессе обучения сети происходит итеративное изменение передаточных весов между входными и выходными узлами сети; тем самым осуществляется поиск оптимального значения критерия группировки. Нейронные сети позволяют эффективно использовать параллельные методы вычислений.

Графовые алгоритмы кластеризации: алгоритм кратчайшего незамкнутого пути и др. Предварительно строится минимальное остовное дерево графа, в котором вершины соответствуют объектам, а ребра имеют длину, равную расстоянию между соответствующими объектами. Для образования кластеров из построенного дерева удаляются ребра максимальной длины.

Иерархический подход: Метод одиночной связи (англ. single linkage, «метод ближайшего соседа»), метод полной связи (англ. complete linkage, «метод дальнего соседа»), метод средней связи (англ. pair-group method using arithmetic averages), Центроидный метод (англ. pair-group method using the centroid average), метод Уорда. Данное направление также имеет отношение к теоретико-графовому подходу. Результаты группировки представляются в

виде дерева группировки (дендрограммы). Алгоритмы, основанные на этом подходе, можно разделить на агломеративные (объединительные) и дивизивные (разделяющие).

Другие методы. Не вошедшие в предыдущие группы, однако широко используемые в кластерном анализе. Статистические алгоритмы кластеризации, ансамбли кластеризаторов, алгоритмы семейства KRAB, DBSCAN (плотностный алгоритм кластеризации пространственных данных с присутствием шума).

Так как результаты кластеризации существенно зависят от выбора метрики, ниже рассмотрены меры расстояний, применяющиеся исследователями в кластерном анализе [5].

Евклидово расстояние. Наиболее широко используемая функция расстояния. Представляет собой геометрическое расстояние в многомерном пространстве. Для точек $p=(p_1, p_2, \dots p_n)$ и $q=(q_1, q_2, \dots q_n)$ евклидово расстояние рассчитывается по формуле (1):

$$d(p,q)=\sqrt{\sum_{k=1}^{n}(p_k-q_k)^2} \quad (1)$$

Квадрат евклидова расстояния. Данная функция расстояния используется в тех случаях, когда необходимо придать больше веса объектам, сильно удаленным друг от друга. Вычисляется квадрат евклидова расстояния так:

$$d(p,q)=\sum_{k=1}^{n}(p_k-q_k)^2 \quad (2)$$

Расстояние городских кварталов (манхэттенское расстояние, расстояние Хэмминга). Это расстояние определяется как среднее разностей по координатам. Во многих случаях эта функция расстояния приводит к

таким же результатам, как и для расстояния Евклида. Функциональное отличие этих метрик состоит в том, что для этой меры влияние отдельных больших разностей (выбросов) уменьшается, так как они не возводятся в квадрат. Хеммингово расстояние вычисляется по формуле (3):

$$d(p,q)=\sum_{k=1}^{n}|p_k-q_k| \quad (3)$$

Расстояние Чебышева. Расстоянием Чебышёва между n -мерными числовыми векторами называется максимум модуля разности компонент этих векторов. Это расстояние используется, когда нужно определить два объекта как «различные», если они различаются по какой-либо одной координате. Рассчитывается по формуле (4):

$$d(p,q)=max\left(|p_k-q_k|\right) \quad (4)$$

Степенное расстояние. Применяется в случае, когда необходимо увеличить или уменьшить вес, относящийся к размерности, для которой соответствующие объекты сильно отличаются. Степенное расстояние вычисляется по следующей формуле:

$$d(p,q)=\sqrt[r]{\sum_{k=1}^{n}(p_k-q_k)^w} \quad , \quad (5)$$

где r и w – параметры, задаваемые субъективно. В случае, когда r и w равны двум, данная метрика совпадет с евклидовой [6].

1.2 Частотные словари

Метод построения частотных словарей (Feature Frequency Profile, FFP) основан на подсчете частот встречаемости олигонуклеотидов (как правило, 2-16 п.н.о.) в последовательности генетического кода [7].

Последовательность X длиной n можно определить как n подряд стоящих символов алфавита A мощности r. Отрезок L символов, где $L \le n$, обозначаем как L *-слово* (L -плет). Тогда множество W_L будет состоять из всех возможных L *-слов*, которые могут быть выделены из последовательности X и иметь K элементов – формула (6).

$$W_L = \left(w_{L,1}, w_{L,2}, \dots w_{L,K}\right) \tag{6}$$

$$K = r^L$$

Иногда поиск L *-слов* может осуществляться с перекрытием вхождений. Вычислительно это осуществляется путем ввода так называемой рамки считывания (скользящего окна), которая считывает последовательность с позиции 1 до $n-L+1$, формула (7).

$$c_L^X = \left(c_{L,1}^X, \dots c_{L,K}^X\right) \tag{7}$$

Для вычисления вектора частот, определяющих последовательность, можно посчитать вероятность $p_{L,i}^X$ нахождения конкретных слов $w_{L,i}^X$, (формула (8)).

$$p_L^X = \left(p_{L,1}^X, p_{L,2}^X, \dots p_{L,K}^X\right) \tag{8}$$

Вектор последовательности f_L^X складывается из частот каждого слова:

$$f_L^X = \frac{c_L^X}{\sum_{j=1}^{K} c_{L,j}^X} \Rightarrow f_{L,i}^X = \frac{c_{L,i}^X}{n-L+1} \qquad (9)$$

Например, для последовательностей ДНК $A = \{\mathrm{T, C, G, A}\}$, $r = 4$, для трехбуквенных слов $L = 3$, возьмем $w_3 = \mathrm{ATC}$. Для короткой последовательности $X = \mathrm{ATATAC}$, где $n = 6$, вектор p_3^X оценивается как отношение частот всех тринуклеотидов. Частоты слов, определяемых сдвигом рамки считывания $n - L + 1 = 4$ раза, будут таковы:

$W_3 = ¿$ {ATA, TAT, TAC, AAA, …}

c_3^X =(2, 1, 1, 0,…)

f_3^X =(0.5, 0.25, 0.25, 0,…)

Векторы c_3^X и f_3^X имеют длину $K = 4^3 = 64$, и нулевые координаты соответствуют пропущенным в X триплетам [8].

На том же примере тринуклеотидов ДНК, можно показать, что, в зависимости от сдвига рамки считывания и позиции стартового нуклеотида, выделяется несколько типов частотных словарей [9].

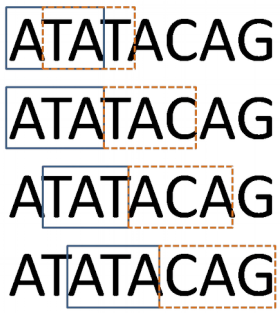


Рисунок 1 – Типы частотных словарей

Первая последовательность – пример построения частотных словарей со сдвигом рамки считывания на 1 нуклеотид, со второй по четвертую – неперекрывающиеся триплеты (со сдвигом рамки считывания на 3 нуклеотида от стартовой позиции) с различными стартовыми позициями.

1.3 Визуализация данных

Задача корректной визуализации данных знакома любому исследователю – к ней сводится проблема представления результатов теоретических исследований или эксперимента в наглядной форме. Также визуализация данных необходима, если в какой-то момент обработки данных возникает желание оценить некоторые характеристики «на глаз».

Визуализация данных – способ представления многомерных данных, в 2- и 3-мерном пространстве, при котором качественно отображены особенности и закономерности распределения данных: кластерная структура, зависимости между признаками, информация о расположении объектов в исходном пространстве [10].

В данной работе задача визуализации данных решалась с помощью программного обеспечения ViDaExpert [11, 12]. ViDaExpert – инструмент визуализации и анализа многомерных векторных данных; в настоящей работе использовались функции построения упругих карт.

Построение упругих карт. Служит обобщением метода главных компонент (в котором вместо упругой пластины используется абсолютно жесткая плоскость). Данный метод был разработан создателями программного обеспечения и основан на алгоритме построения упругой сети.

Упругая сеть – связанный и упорядоченный граф G(Y, E), где $Y=\{y^{(i)}, i=1\ldots r\}$ обозначает множество узлов графа, $E=\{E^{(i)}, i=1\ldots s\}$ –

множество ребер графа. Объединим некоторые смежные ребра в пары $R^{(i)}=\{E^{(i)},E^{(k)}\}$ и обозначим через $R=\{R^{(i)},i=1\dots r\}$ множество ребер жесткости графа.

У каждого ребра $E^{(i)}$ есть начальный узел $E^{(i)}(0)$ и конечный узел $E^{(i)}(1)$. Ребро жесткости – пара смежных ребер, имеет начальный $R^{(i)}(1)$ и конечный $R^{(i)}(2)$ узлы, а также центральное ребро $R^{(i)}(0)$.

Рисунок 2 – Узел, ребро, ребро жесткости

Узлы сетки располагаются в многомерном пространстве данных. Это может быть сделано несколькими способами: располагая узлы случайно или в выделенном подпространстве. Например, граф может быть расположен на линейном многообразии, натянутом на первые две или три главные компоненты. В любом случае каждый узел графа становится вектором R^M , где M – выбранная мера расстояния.

Функция энергии U, в которой просуммированы вклады каждого узла, ребра и ребра жесткости определяется так:

$$U=U^{(Y)}+U^{(E)}+U^{(R)} \qquad (10)$$

Разделим все множество точек данных на подмножества $K^i,i=1\dots p$. Каждое из подмножеств содержит объекты, для которых $y^{(i)}$ является ближайшим узлом:

$$K_i = \left\{ x^{(j)} : \left\| x^{(j)} - y^{(i)} \right\| \vec{Y} \, min \right\} \quad . \qquad (11)$$

Определим

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

$$U^{(E)} = \sum_{i=1}^{s} \lambda_i \left\| E^{(i)}(1) - E^{(i)}(0) \right\|^2 \quad , \qquad (12)$$

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

Величина $U^{(Y)}$ является средним квадратом расстояния между узлом сетки $y^{(i)}$ и точками данных в $K^{(i)}$, $U^{(E)}$ – аналог суммарной энергии упругого растяжения сетки и $U^{(R)}$ – аналог суммарной энергии упругой деформации сетки.

Упругая сеть строится с помощью минимизации суммарной энергии U графа G. В евклидовом пространстве процедура минимизации может быть реализована итерационным алгоритмом, на каждом шаге которого производится разбиение множества точек данных на подмножества по критерию близости к узлам графа и выполняется процедура квадратичной оптимизации, в результате чего узлы графа принимают новые положения в пространстве данных.

После того, как упругая сеть построена, она может выступать основой для построения упругой карты – нелинейного аппроксимирующего многообразия.

Визуализация данных осуществляется с помощью проецирования точек данных на построенную упругую карту, что позволяет показывать проекции точек данных в пространстве малой размерности.

1.4 Метод BLAST (Basic Local Alignment Search Tool)

Данный алгоритм широко применяется для поиска гомологов нуклеиновых кислот и белков с частично или полностью известной первичной структурой [13]. Используя BLAST, можно сравнить имеющуюся последовательность с последовательностями баз данных (GenBank, UniProt EMBL DDBJ PDB RefSeq) и, исходя из степени гомологичности, сделать некоторые выводы о структуре и функциях последовательности.

В основе BLAST лежат надежные алгоритмы выявления близкородственных участков разных последовательностей путем локального попарного выравнивания (ускоренный почти в 50 раз аналог алгоритма Смита-Ватермана) и функции статистической обработки результатов, которые обусловили ее повсеместное использование в научной практике [14].

2 Материалы и методы

2.1 Характеристики сборки и получение выборок

Для работы были использованы результаты ассемблирования генома лиственницы сибирской. Данные геномного секвенирования лиственницы сибирской были получены в лаборатории геномных исследований СФУ под руководством проф. К.В. Крутовского.

Секвенирование осуществлялось на приборе Illumina. Все данные получены из одного мегагаметофита. Сборка осуществлялась специалистами лаборатории геномных исследований на высокопроизводительном сервере IBM x3950.

Таблица 1 – Характеристики используемой сборки

Показатель	Значение
Число контигов	4591622
Максимальная длина контига	385868
Общая длина	5586881443
N50	1949
N90	514

Для дальнейшего анализа нами были выбраны несколько групп контигов. Это обусловлено тем, что полученная сборка имела достаточно большой размер и использование всего объема данных существенно бы замедлило и усложнило работу. Контиги в сборке имели достаточно большой разброс по длинам (от 200 до 385868 п.н.о.) и выборки были составлены таким образом, чтобы охватить весь ряд. Однако же в самих выборках колебания длин, как правило, незначительны.

Таблица 2 – Геномные выборки

№	Длина контигов в группе, п.н.о.	Число контигов
1	15 000 – 385 868	4 271

2	10 000	4772
3	3 000	5338
4	723	2503
5	200	8 545

Далее были построены частотные словари для каждого контига в выборке. В данной работе рассматривались частоты встречаемости трехбуквенных нуклеотидов со сдвигом рамки считывания в один нуклеотид. Для получения частотных словарей был использован скрипт, написанный сотрудниками лаборатории геномных исследований на языке программирования Perl.

Для дальнейших исследований из частотных словарей исключался один тринуклеотид, так как сумма частот всех триплетов в словаре равна единице. Исключался триплет с минимальным стандартным отклонением.

2.2 Метод динамических ядер (k-means)

Для кластеризации нами был выбран метод динамических ядер – простейший базовый алгоритм кластерного анализа, основанный на вычислении центров масс и минимизации суммарного квадратичного отклонения точек кластеров от центров. Алгоритм включает в себя несколько последовательных шагов:

1. Случайно выбрать k точек и обозначить их начальными центрами масс кластеров;
2. Распределить объекты по кластерам с ближайшим центром масс;
3. Пересчитать центры масс кластеров согласно текущему положению и посчитать суммарное квадратичное отклонение точек кластеров от центров этих кластеров;

$$V=\sum_{i=1}^{k}\sum_{x_j\in S_i}\left(x_j-\mu_i\right)^2 \quad , \qquad (13)$$

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

4. Если условия остановки алгоритма не выполнены, вернуться к пункту 2 [15].

Условия остановки алгоритма: кластерные центры стабилизировались, все объекты принадлежат кластеру, которому принадлежали до текущей итерации; число итераций достигло максимального числа итераций.

К недостаткам и проблемам алгоритма можно отнести то, что не гарантируется достижение глобального минимума суммарного квадратичного отклонения V, а только одного из локальных минимумов. Также результат сильно зависит от выбора исходных центров кластеров, их оптимальный выбор неизвестен.

Сложной задачей для исследователя является выбор числа кластеров. Если нет предположений относительно этого числа, рекомендуют создать 2 кластера, затем 3, 4, 5 и т.д., постоянно сравнивая полученные результаты. [16]

К достоинствам метода динамических ядер относятся: простота использования и быстрота работы, понятность и прозрачность алгоритма. Также достаточно легко оценить разделимость кластеров. Для этого рассчитываются средние значения для каждого кластера. При хорошей кластеризации средние кластеров будут существенно отличаться. Более подробно с алгоритмом можно ознакомиться в работах [17, 18].

2.3 Второе обобщенное правило Чаргаффа и величина невязки

Согласно гипотезе, выдвинутой в работе [9], при кластеризации нуклеотидных последовательностей на 2 группы, существует возможность разделение данных по принадлежности к стренду ДНК. Проверить эту гипотезу можно достаточно простым способом, основанным на втором правиле Чаргаффа. Второе правило Чаргаффа гласит: в одном стренде ДНК

количество тимина (T) примерно равно количеству аденина (A), а количество цитозина (C) – количеству гуанина (G). Точность этого равенства постепенно падает с увеличением размера исследуемой последовательности. Второе правило универсально – ему строго подчинены все прокариоты и эукариоты, а также вирусы, содержащие двунитевую ДНК (dsDNA).

С целью проверки кластеризации рассчитывается величина невязки, описывающая меру отклонения кластера от вышеупомянутого равенства [19]. Для расчетов берутся частоты тринуклеотидов и их комплиментарных палиндромов (читающихся одинаково в противоположных направлениях, с учетом замены по правилу комплиментарности). Невязка вычисляется по формуле:

$$\mu = \frac{1}{\omega}\sqrt{\sum \left(f_w - f_{\acute{w}}\right)^2} \quad , \qquad (14)$$

где ω – количество комплиментарных пар, f_w – частота встречаемости триплета и его комплиментарного палиндрома $f_{\acute{w}}$.

3 Результаты и обсуждение

3.1 Выборка самых длинных контигов

Основываясь на характеристиках частотных словарей, мы предположили, что на больших длинах метод покажет наилучшие результаты. А так как первая выборка имеет достаточно большой разброс по длинам (15 000 – 385868 п.н.о.), ожидалось увидеть хорошую разделимость кластеров.

После построения упругой карты и отображения ее во внутренних координатах (рисунок 3) мы заметили, что выборка отлично делится, как минимум, на три класса. Можно схематично выделить кластеры на картинке, позднее мы к ним вернемся.

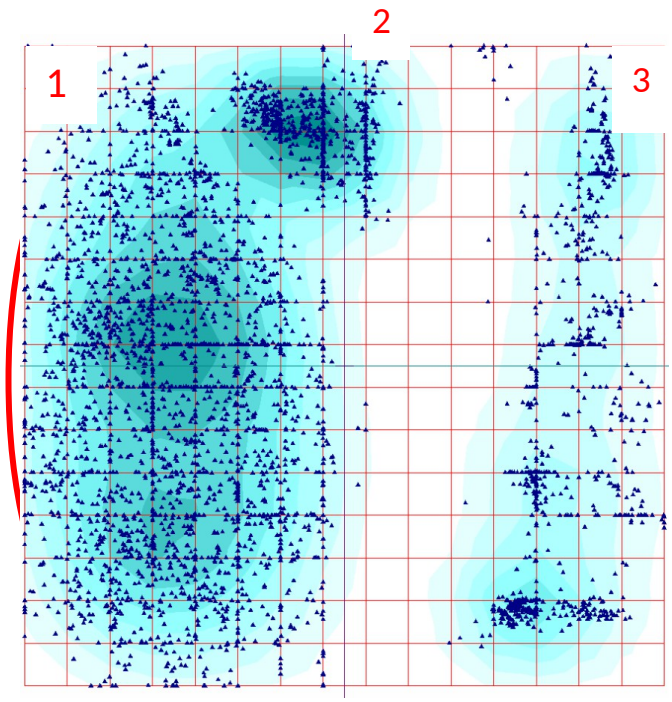


Рисунок 3 – Отображение распределения контигов первой выборки во внутренних координатах карты

Кластерный анализ выборки начинали с деления на 2 кластера. Деление оказалось весьма устойчивым – в 98 случаях из 100 контиги распределялись

между кластерами определенным способом. Для проверки результатов были посчитаны радиусы кластеров и расстояние между их ядрами (рисунок 4).

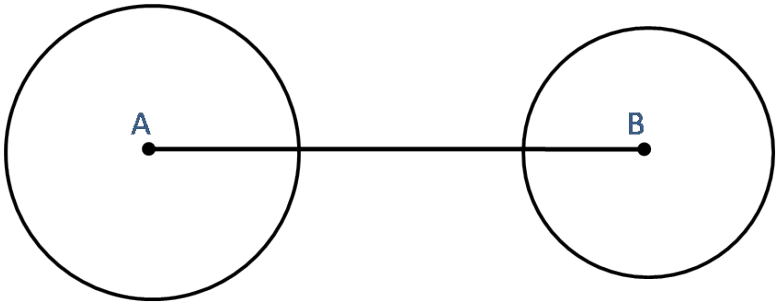


Рисунок 4 – Схематическое отображение радиусов кластеров и расстояние между ними

Далее были посчитаны невязки внутри классов: 0,0000255 и 0,0000171; между классами: 0,00146. То, что величина невязки внутри классов примерно на два порядка меньше, чем между ними, позволяет предположить, что для этой выборки, в каждый класс попали контиги, принадлежащие обоим стрендам – результат, противоположный распределению контигов при кластеризации транскриптомных данных.

Далее было произведено деление на 3 класса. Контиги кластеризовались устойчиво, один из кластеров, полученных при делении на 2, сохранил свою целостность и при всех последующих делениях. Деление на 4 класса оказалось неустойчивым и не несло информативных результатов. Ниже схематично отображено расхождение контигов по кластерам (рисунок 5).

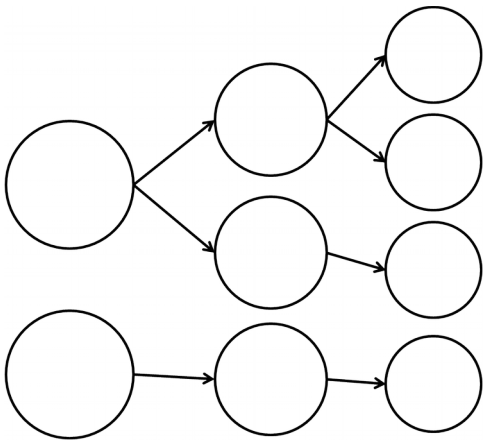


Рисунок 5 – Схема расхождения контигов при кластеризации

Составы кластеров, полученных при разбиении на 3 (выделены и пронумерованы на рисунке 3) были проанализированы с помощью сервиса BLAST.

Выравнивание контигов первого класса на нуклеотидную базу показало следующие результаты: была обнаружена высокая степень гомологичности с последовательностями ядерного генома хвойных, таких, как лиственница европейская, лиственница Гмелина, лиственница западная, пихта белая европейская, пихта великая, кедр гималайский, ель обыкновенная, ель канадская, сосна кедровая европейская, сосна ладанная и других. Также было найдено некоторое сходство контигов с последовательностями геномов других растений, например, аробидопсиса, люцерны и томата.

Кластер 2 показал высокое сходство с участками митохондриальных и хлоропластных геномов хвойных и других растений. Мы предположили, что в этот кластер объединились последовательности геномов органелл.

Третий кластер показал самые спорные результаты выравнивания. Согласно сервису BLAST, эта группа последовательностей имеет высокую степень гомологичности с последовательностями геномов бактерий. Самый высокий уровень идентичности с *Escherichia coli, Enterobacter sp, Pseudomonas chlororaphis, Klebsiella pneumoniae, Xanthomonas citri.*

Причин у этого может быть несколько. Во-первых, высокую идентичность нуклеотидных последовательностей у разных видов может вызывать наличие высококонсервативных некодирующих последовательностей ДНК, а также консервативные гены. Во-вторых, нельзя исключать возможность привнесения бактериальной ДНК на одном из этапов пробоподготовки и секвенирования, т.е. бактериальной контаминации. Так или иначе, необходим дальнейший анализ третьего кластера.

3.2 Выборка контигов, длиной 10000 п.н.о.

Затем была построена упругая карта для контигов второй выборки (длиной 10 000 – 11 120 п.н.о.). На рисунке видно, что разделение на группы менее очевидно по сравнению с первой выборкой.

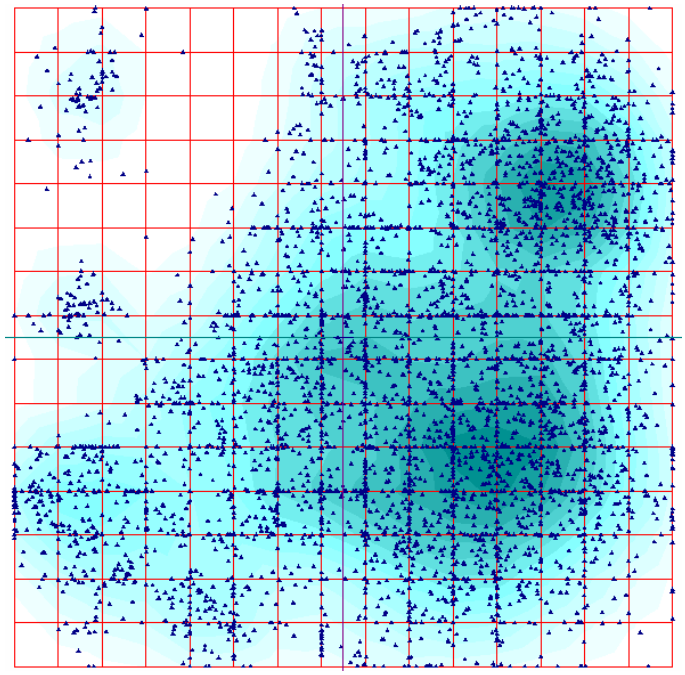


Рисунок 6 – Отображение распределения контигов второй выборки во внутренних координатах карты

При делении выборки на две группы в 62 случаях из 100 контиги контиги распределялись по группам одинаково. Однако в этих же 62 случаях,

одна из групп была представлена одним контигом. При детальном рассмотрении оказалось, что этот контиг содержит длинный концевой повтор, что сильно отражается на частотах определенных триплетов, и следовательно, делает этот контиг отличным от других.

Далее мы делили выборку на 3 группы. Деление сохранило 62% устойчивость, третий кластер – отделившиеся от общей группы 72 контига. Ниже представлена разделимость трех кластеров.

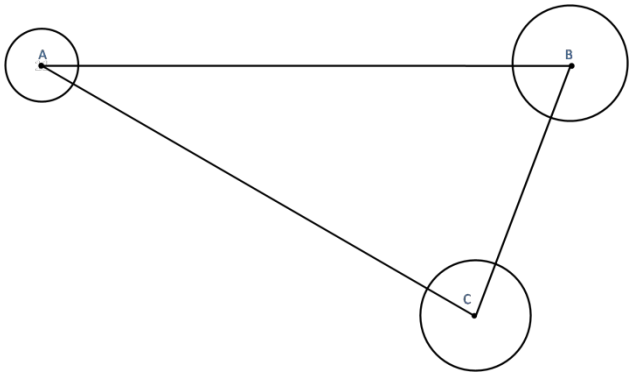


Рисунок7 – Схематическое отображение радиусов кластеров и расстояния между ними

Проверка третьего кластера с помощью BLAST показала схожесть принадлежащих этому кластеру контигов с последовательностями бактериальных геномов.

3.3 Выборка контигов, длиной 3000 п.н.о.

Третья выборка состояла из контигов длиной 3 000 – 3 025 п.н.о. и при визуализации выглядела вот так:

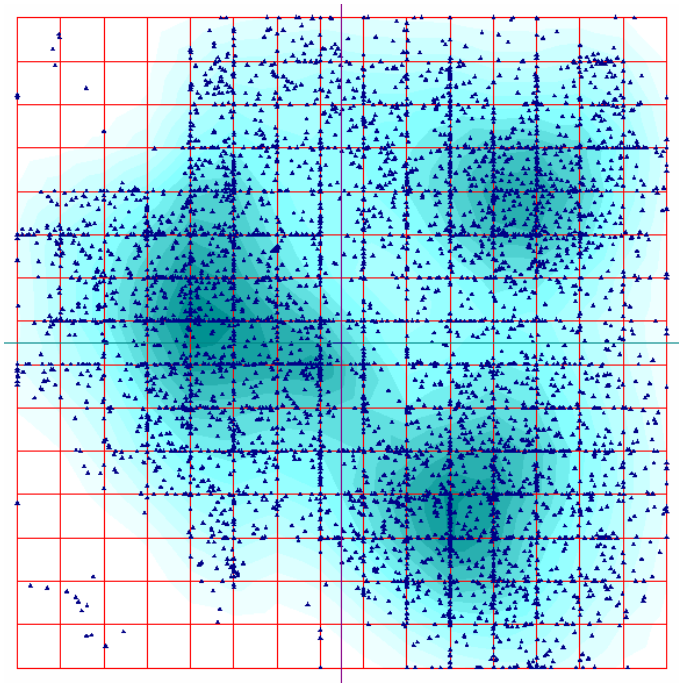


Рисунок 8 – Отображение распределения контигов третьей выборки во внутренних координатах карты

Результаты кластеризации этой выборки оказались крайне неустойчивыми, но нас заинтересовала ситуация, когда при делении на два класса в отдельную группу объединялись то 13 контигов в левом верхнем углу карты, то 14 в левом нижнем. Деление на 3 класса оказалось чуть более устойчивым (21 случай из 100), а полученные классы обладали высокой разделимостью (рисунок 9).

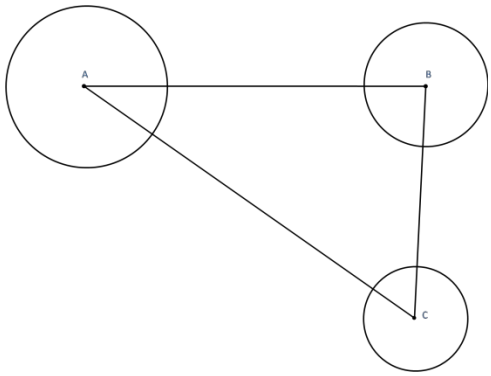


Рисунок 9 – Схематическое отображение радиусов кластеров и расстояния между ними

Представители группы 13 контигов оказались высоко гомологичны последовательностям геномов различных видов рода *Pseudomonas* и некоторых других видов бактерий. Проверка в BLAST кластера, состоящего из 14 контигов, оказалась не результативной. Мы установили, что эти контиги богаты моно- и олигонуклеотидными тандемными повторами, и тот факт, что они обособились в одну группу, обусловлен отличием частот встречаемости высокоповторяющихся триплетов.

3.4 Выборка длин по среднему значению

Результаты визуализации выборки 4, в которую входят контиги длиной по 723 п.н.о. – среднее значение в используемой сборке. На рисунке 10 можно заметить, что некоторые контиги выстроились по линиям координатной сетки. Это обусловлено тем, что на такой длине частоты многих триплетов равны, некоторые триплеты не встречаются вовсе.

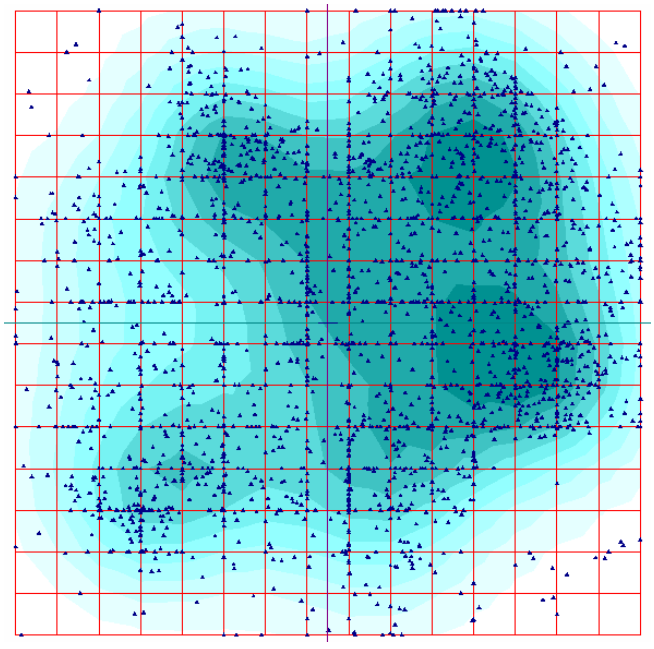


Рисунок 10 – Отображение распределения контигов четвертой выборки во внутренних координатах карты

При разбиении выборки на две группы, в 54 случаях из 100 контиги распределяются в группы одинаково. Однако кластеры при таком распределении обладают низкой разделимостью (рисунок 11). При дальнейшем делении на 3 и 4 класса картина становится крайне неустойчивой, и для дальнейшего анализа нами были выбраны результаты кластеризации на 2 группы.

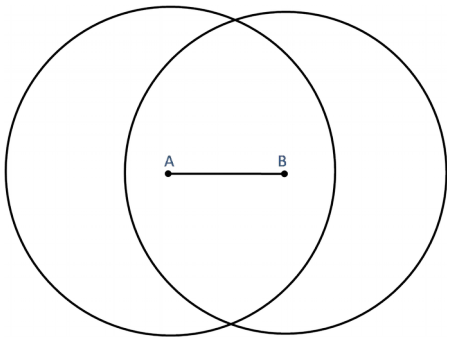


Рисунок 11 – Схематическое отображение радиусов кластеров и расстояния между ними

Далее были посчитаны невязки внутри классов: 0,000081 и 0 0,000066; между классами: 0,000735. Как и для первой выборки, величина невязки внутри классов оказалась существенно меньше таковой между ними, то есть, контиги из разных стрендов ДНК распределились по обеим группам.

3.5 Выборка наименьших длин

Работа с выборкой самых маленьких (200 п.н.о.) контигов начиналась с построения упругой карты (рисунок 12). Как видно на рисунке 12, эффект распределения контигов по линиям координатной сетки еще более выражен, по сравнению с предыдущей выборкой.

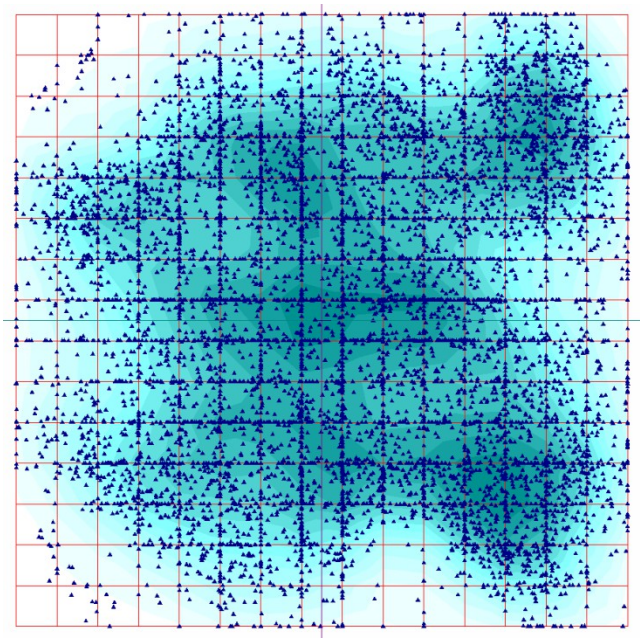


Рисунок 12 – Отображение распределения контигов пятой выборки во внутренних координатах карты

При делении выборки на два класса кластеры неустойчивы и обладают низкой разделимостью (рисунок 13). Вероятно, это также обусловлено потерей чувствительности частотных словарей на коротких длинах.

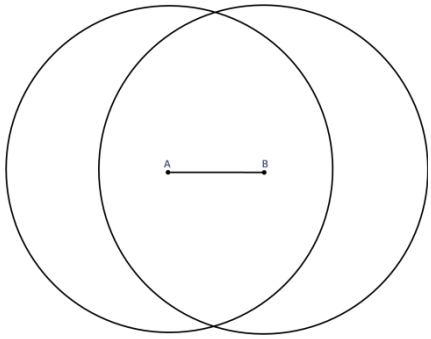


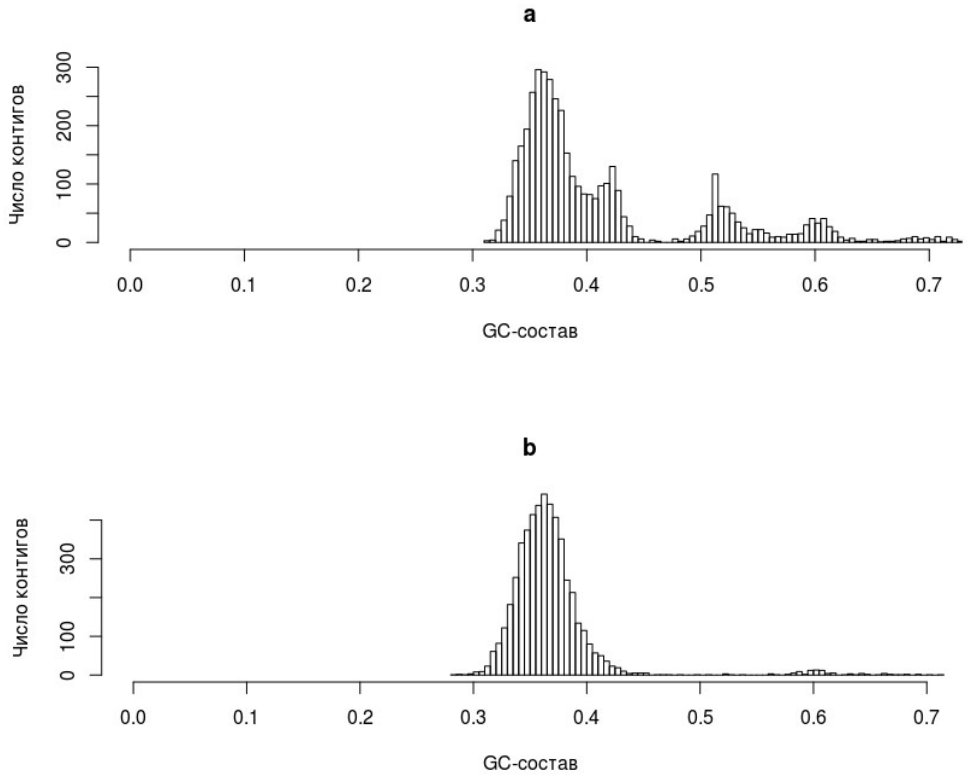
Рисунок 13 – Схематическое отображение радиусов кластеров и расстояния между ними

Тем не менее, нами были посчитаны невязки внутри классов: 0,000072 и 0,000076, между классами: 0,001712. И опять величина невязки между классами оказалась гораздо больше, чем внутри каждого класса. Почему такое отличие от результатов кластеризации транскриптомной сборки? Является ли разделение по стрендам характеристикой, присущей транскриптому, или же особенности нашей сборки (например, наличие митохондриальных последовательностей) влияют на распределение?

3.6 GC-контент выборок

GC-контент (GC-состав) – суммарная доля всех гуанинов (G) и цитозинов (C) по отношению к длине исследуемого участка нуклеиновых кислот. Для каждого контига в выборке был посчитан GC-контент, распределение в выборках показано на рисунке 14.

Согласно данным, ранее полученным в лаборатории геномных исследований, GC-контент для сборки ядерного генома лиственницы сибирской, составляет в среднем 38%. На гистограммах видно, что высокое соответствие этим данным показывают контиги второй (b) и третей (c) выборок.



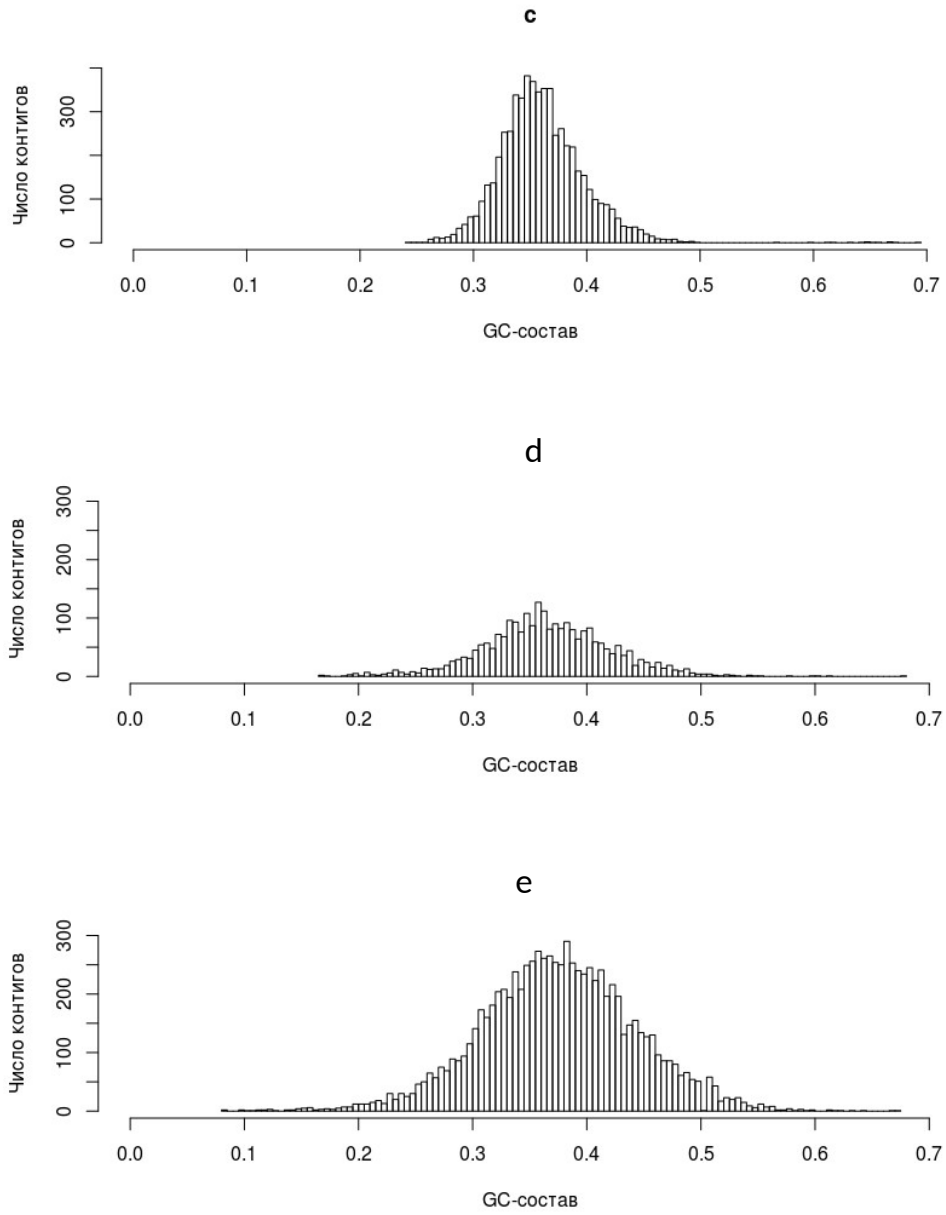


Рисунок 14 – Распределение GC-контента с первой выборки (a) по пятую (e)

Таким образом, сравнивая данные из таблицы 3 с эталонным значением, видно, что в первой выборке присутствуют последовательности, сильно отличающиеся по GC-контенту. Остальные выборки более-менее укладываются в распределение.

Таблица 3 – GC-контент

№ выборки	Среднее	Стандартное отклонение
1	0.4141326	0.08521772
2	0.3677757	0.04150827
3	0.3608543	0.03793649
4	0.3650961	0.05650409
5	0.375792	0.0681856

Пологий пик распределения GC-контента в пятой выборке (e) обусловлен тем, что, как правило, контиги длиной 200 п.н.о. имеют самое низкое качество. А если добавить к этому то, что кластеризация частотных словарей для таких длин оказалась не результативной, в дальнейшем планируется отказаться от работы с выборками минимальных длин.

ЗАКЛЮЧЕНИЕ

В данной работе показано, что кластеризация методом динамических ядер позволяет разделить последовательности генома на структурно различные группы. Подтверждено снижение чувствительности частотных словарей по мере уменьшения длин последовательностей, составляющих выборку.

Геномная сборка лиственницы не однородна по структурному составу, встречаются последовательности, резко отличающиеся по частотам триплетов и GC-контенту. Возможное объяснение такой разнородности – примеси митохондриального генома и бактериальные контаминации. Необходимо продолжить анализ таких последовательностей.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Nystedt B, Street NR, Wetterbom A et al. The Norway spruce genome sequence and conifer genome evolution // Nature. – 2013. – 497(7451);
2. Мандель, И.Д. Кластерный анализ / Финансы и статистика. – Москва, 1988. – С. 10;
3. Воронцов, К.В. Алгоритмы кластеризации // Лекции по алгоритмам кластеризации и многомерного шкалирования. – 2007. – С. 2;
4. Бериков, В.Б., Лбов, Г.С. Современные тенденции в кластерном анализе // Информационно-телекоммуникационные системы. – 2008. – С. 6–9;
5. Ichino M., Yaguchi H. Generalized Minkowski Metrics for Mixed Feature-Type Data Analysis // Advances in Data Science and Classification. – 1994. – Pp. 698–708;
6. Deza E., Deza M.M.. Encyclopedia of Distances / Springer-Verlag Berlin Heidelberg, 2009. – Pp. 583;
7. Sims GE, Jun S-R, Wu GA, Kim S-H. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions // Proceedings of the National Academy of Sciences of the United States of America. – 2009. – 106(8). Pp. – 2677–2682;
8. Vinga S, Almeida J. Alignment-free sequence comparison—a review // Bioinformatics. – 2003. – 19(4). Pp. – 513–523;
9. Sadovsky, M.G., Birukov, V.V., Putintseva, Y.A., Oreshkova, N.V., Vaganov, E.A. and Krutovsky, K.V. Symmetry of Siberian Larch Transcriptome // Journal of Siberian Federal University: Biology. – 8(3). – Pp. 278-286.
10. Зиновьев, А.Ю. Визуализация многомерных данных / Изд. КГТУ. – Красноярск. – 2000. С. – 180.
11. Gorban A, Pitenko A, Zinovyev A. ViDaExpert: user-friendly tool for nonlinear visualization and analysis of multidimensional vectorial data // arXiv:1406.5550v2. – 2014. P. – 9.
12. ViDaExpert - is a software for multidimensional vectorial data visualization http://bioinfo-out.curie.fr/projects/vidaexpert/;
13. Altschul S., Gish W., Miller W., Myers E., and Lipman D. Basic local alignment search tool // Journal of Molecular Biology. – 1990. – 215(3);
14. Madden TL, McGinnis S. Blast: at the core of a powerful and diverse set of sequence analysis tools // Nucleic Acids Res. – 2004;

15. Котов А., Красильников Н. Кластеризация данных [Электронный ресурс]. – 2006. С. – 16;
16. Нейский И.М. Классификация и сравнение методов кластеризации // Интеллектуальные технологии и системы. Сборник учебно-методических работ и статей аспирантов и студентов. – М. – 2006. – С. 130–142;
17. Telgarsky M., Vattani A. Hartigan's Method: k-means Clustering without Voronoi // Journal of Machine Learning Research. – 2010. P. – 9;
18. Hartigan JA. Algorithm AS 136: A K-Means Clustering Algorithm // Journal of the Royal Statistical Society. – 1979. – 28(1). Pp. – 100–108;
19. Гребнев, Я. В., Садовский, М. Г. Второе правило Чаргаффа и симметрия геномов // Фундаментальные исследования. – 2014. №12-5. С. – 4.