


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

Институт математики и фундаментальной информатики

Базовая кафедра вычислительных и информационных технологий

УТВЕРЖДАЮ

Заведующий кафедрой

 В.В.Шайдуров

« ____ » _____ 2017 г.

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

**ИССЛЕДОВАНИЕ СТАТИСТИЧЕСКИХ СВОЙСТВ СИМВОЛЬНЫХ
ПОСЛЕДОВАТЕЛЬНОСТЕЙ ПО НАБОРАМ НАИКРАТЧАЙШИХ
«ПОТЕРЯННЫХ» СТРОК**

Направление 02.04.01 Математика и компьютерные науки

Магистерская программа 02.04.01.02 Вычислительная математика

Научный руководитель

к.ф.-м.н., доцент



М.П. Варыгина

Выпускник

Ю.О. Якубайлик

Красноярск 2017

Аннотация

В генетике с точки зрения математика ДНК – это последовательность нуклеотидов, представленных в виде символьной последовательности.

Символы представляют азотистые основания. Была написана программа поиска символьных последовательностей, отсутствующих в ДНК. Обработано более 400 нуклеотидных последовательностей, представляющих геномы организмов разного таксономического положения: от бактерий до человека.

Ключевые слова: потерянная последовательность ДНК, таксономия.

In genetics, mathematics thinks that DNA is a sequence of nucleotides, represented as a character sequence. Characters represent the nitrogen bases. A program developed to search character sequences that avoided in DNA. Processed more than 400 nucleotide sequences representing the genomes of organisms of different taxonomic position: from bacteria to human.

Keywords: the lost DNA sequence, taxonomy.

СОДЕРЖАНИЕ

Введение.....	4
Глава 1.....	6
Глава 2.....	8
Глава 3.....	10
Список использованных источников.....	36
Приложение А.....	37

ВВЕДЕНИЕ

В генетике существует три точки зрения на понятие ДНК. С биологической точки зрения ДНК представляет собой макромолекулу, обеспечивающую хранение, передачу из поколения в поколение и реализацию генетической программы развития и функционирования живых организмов. С химической точки зрения ДНК — это длинная полимерная молекула, состоящая из повторяющихся блоков — нуклеотидов. Каждый нуклеотид состоит из азотистого основания, сахара (дезоксирибозы) и фосфатной группы. В ДНК встречается четыре вида азотистых оснований (аденин, гуанин, тимин и цитозин). Азотистые основания одной из цепей соединены с азотистыми основаниями другой цепи водородными связями согласно принципу комплементарности: аденин соединяется только с тимином, гуанин — только с цитозином. С точки зрения математика ДНК суть последовательность нуклеотидов, представленных в виде символической последовательности. Символы представляют азотистые основания одной из цепей.

Задач, связанных с символическими последовательностями, можно придумать очень много, например, выделение упорядоченности, подсчет статистики различных комбинаций символов, и т.п. Тем не менее, в этой работе будет освещено в частном порядке именно анализ ДНК. Какую бы задачу мы ни решали в связи с анализом ДНК, в конечном итоге она почти наверняка сводится к изучению частотного словаря. Носитель частотного словаря — это все символические последовательности определенного количества символов; в данной работе назовем их словами. Если каждому такому слову сопоставить натуральное число его копий включая нуль, получится конечный словарь. А если число копий заменить на частоту их появления, то есть число копий данного слова, деленное на сумму чисел копий всех слов, то получится частотный словарь. Рано или поздно наверняка какие-то достаточно длинные слова из носителя пропадут, то есть число копий слов будет равно нулю. Назовем их «потерянными» словами.

Целью данной работы было выявить связи вышеназванных «потерянных» слов с таксономией, то есть положением в живом мире, и другими биологическими показателями. Предмет исследования включает в себя списки «потерянных» слов и их свойства: насколько состав списков связан с конкретными организмами (или группой организмов) и насколько он связан со специфическими (короткими) последовательностями, играющими важную регуляторную роль. Объектом являются комбинаторные и статистические свойства таких «потерянных» слов. То есть рассматривается состав слов, «потерянных» определенными последовательностями, и их таксономия.

Задачами магистерской работы были:

- написание программы, вычисляющую список «потерянных» слов,
- подбор релевантного генетического материала (геномы подходящих организмов),
- обработка этого материала, анализ полученных списков слов (их подобие у разных организмов, связь с GC-контентом и прочее) и выявление связи установленных списков потерянных слов с таксономией носителей соответствующих последовательностей.

Глава 1

Анализ частотных и статистических свойств нуклеотидных последовательностей является областью, исследований, в которой работает большое число исследователей.

Кутыркин В.А. и Чалей М.Б. ввели понятие скрытой профильной периодичности в последовательностях ДНК. В 2013 году была опубликована статья, в которой был изложен метод распознавания такого типа скрытой периодичности [1]. Этот метод основан на применении спектрально-статистического подхода (2С-подхода).

Мирошниченко Л.А. и Гусев В.Д. в 2014 году написали программу [2], предназначенную для выявления фрактальных и фракталоподобных структур в ДНК-последовательностях. Ядром таких структур являются палиндромы или комплементарные палиндромы. Программа может быть использована для анализа структуры и классификации ДНК-последовательностей. Программа написана на языке Pascal в программной среде Delphi 7.

Спангенберг В.Е., Дадашев С.Я., Коломиец О.Л., Богданов Ю.Ф. в 2011 году исследовали [3] методами компьютерного анализа связь распределения повторяющихся последовательностей (ПП) с распределением событий мейотической рекомбинации в хромосоме 1 мыши и методом гибридизации *in situ* расположение разных типов ПП относительно синаптонемного комплекса на препаратах (спредах) сперматоцитов мыши.

Но в данных публикациях нигде не упоминаются списки потерянных последовательностей ДНК. Поэтому лучше обратиться к зарубежным источникам.

Арун Рават и Юйпин Денг в 2008 году разработали метод [4], основанный на теории графов, который учитывает предположение биологии о том, что для условной сорегуляции необходимо добывать большие банки транскрипционных данных и свойства данных микромассивов. Алгоритм вычисляет взаимосвязь между генами путем преобразования дискретных

сигналов из временных рядов данных микромассивов (технология AtGenExpress).

Бейлин Хао, Хуймин Ксие, Зугуо Ю и Гуойи Чен в 1998 году задались [5] визуализацией потерянных либо недопредставленных строк в некоторых бактериальных геномах двумя способами: кластеризацией методом Гоулдена-Джексона или конструированием минимального конечного автомата, определенного набором запрещенных слов соответствующего языка.

Апробация работы была описана в испанском журнале на английском языке [6].

Задача состоит в выявлении связей потерянных слов с таксономией. Для этого нужно выделить сами потерянные слова. Это делается программой, написанной на языке C++ в среде программирования MSVisualC++ 2008. Она также преобразует файлы в форматы, удобные для рассмотрения. Листинг можно найти в приложении. Также можно выделить задачу обработки материала. Сколько времени займет программная обработка, зависит от длины минимального потерянного слова. Отметим, что это последовательная версия программы.

Глава 2

Материалы брали в EMBL-банке; традиционно, там депонируют самые хорошо секвенированные, хорошо асSEMBЛИРОВАННЫЕ и хорошо аннотированные геномы. Брали представителей разных таксономических групп. Структура этих файлов такова, что первые строки занимает заголовок, а сам геном поделен строками, в которых 6 блоков по 10 значащих символов.

Для использования таких данных надо отсечь заголовки и объединить геном в одну большую строку. После этого, если строка слишком велика, разбиваем ее на части с перекрытием, чтобы не пропустить слова на конце-начале.

Так как мы не знаем, какие слова пропали, надо составить носитель словаря, то есть все возможные комбинации, их 4^K , где K – длина слова. Также если список длинный, делим его на части.

Для каждого файла проверяем слова каждой длины, ищем первое пропавшее слово, длина которого и будет для данного файла наименьшей длиной пропавших слов.

Проверка идет по каждому куску файла по каждому куску носителя словаря и по каждому слову. Если есть слово, записываем, что в файле оно присутствует и забываем про него. Если нет, то это тоже записываем, но запоминаем, что в этом куске файла слова нет, и переходим к следующему слову. Если кончились слова, переходим к следующему куску носителя. Если кончился носитель, переходим к следующему куску файла. Если кончился файл, то оставшиеся слова и будут нужными словами.

Программа преобразует файлы в форматы, удобные для рассмотрения, и делает это после проверки. К тому же, эту часть программы можно запустить отдельно. Основной формат – просто строка с пробелами – разделителями слов.

Формат стека с файлами появляется, когда есть папка с файлами в основном формате: выписываются все слова определенной длины, которые

принадлежат папке, все файлы, которым принадлежит конкретное слово, и количество таких файлов. И так для всех длин в папке.

Табличный формат, как следует из названия – таблица, в которой столбцы обозначают файлы, строки – слова, а в самой таблице ставится плюс, когда файлу принадлежит слово, и минус – когда не принадлежит. Данный формат реализован для слов, принадлежащих двум и более файлам.

Формат совпадений с меньшей длиной слов появляется, когда есть соседние файлы в плане длин слов формата стека с файлами. Программа проверяет файл с большим словом на наличие в нем слов из файла с меньшим словом, записывая оба, если оно есть.

При увеличении длины слова происходит показательный рост полной длины носителя.

$$q = 4^l$$

где q – полная длина носителя, l – длина слова.

Полная длина носителя длины минимального потерянного слова сопоставима с длиной последовательности. Следовательно, разумно наблюдать крайне медленный рост длины потерянных слов при увеличении длины последовательности. Но похоже, что длина потерянного слова растёт ещё медленнее. Приведённая оценка – оценка сверху.

Для каждой последовательности мы можем вычислить её GC-контент – общее число букв G и C, делённое на общую длину последовательности.

Глава 3

Обработано более 400 нуклеотидных последовательностей, представляющих геномы организмов разного таксономического положения: от бактерий до человека.

О времени выполнения надо сказать отдельно. На 9-10 символах на один файл уходит полтора часа, в то время, как на 6-7 символах за минуту программа считает примерно 50 файлов.

Установлено, что внутригеномная вариабельность списков очень велика: списки потерянных слов, определяемых для разных хромосом одного генома имеют очень малые пересечения. Это означает, что с точки зрения потерянных слов мы должны рассматривать весь геном как одно целое: если мы все хромосомы соединим в одну длинную большую последовательность, то список потерянных слов почти не изменится.

Полный список потерянных слов минимальной длины вида *Anopheles gambiae* приведен в формате стека с файлами в таблице 1.

Таблица 1 – *Anopheles gambiae*, длина потерянных слов – 10.

Слово	Файлы	Слово	Файлы	Слово	Файлы	Слово	Файлы
aaccatagg	cm000360;	ccctcatagg	cm000360;	gacccttaga	cm000360;	gtaaccgcct	cm000360;
aacctgggat	cm000360;	ccctgactag	cm000360;	gacctactca	cm000360;	gtaagaccac	cm000360;
aagactggtc	cm000360;	ccgactatac	cm000360;	gacctagtcc	cm000360;	gtaaggcctg	cm000360;
aaggccctac	cm000360;	ccgacttagg	cm000357;	gacctatcac	cm000360;	gtaccagtc	cm000360;
aataccagtc	cm000360;	ccgagtctct	cm000360;	gaccttagat	cm000359;	gtactcaggg	cm000360;
aatcctgagg	cm000360;	ccgcatagta	cm000360;	gaccttagtc	cm000360;	gtagactcaa	cm000360;
acagtctcta	cm000360;	ccgcgactat	cm000360;	gacgcctgat	cm000360;	gtagggagtc	cm000360;
acataggccg	cm000360;	ccgctaagag	cm000360;	gacgcctgag	cm000356;	gtagggggcc	cm000360;
accataggcg	cm000360;	ccgctgagcc	cm000360;	gacgtctcac	cm000360;	gtagggtacc	cm000360;
accatagggg	cm000360;	ccgggtaac	cm000360;	gactaacggt	cm000360;	gtagggtccc	cm000360;
accagacat	cm000360;	ccgtctaaga	cm000360;	gactaagtc	cm000360;	gtagtcaggc	cm000360;
accatagat	cm000360;	cctaagagtg	cm000360;	gactacgtat	cm000360;	gtagctctca	cm000360;
accctagggt	cm000360;	cctaagggtc	cm000360;	gactagaggt	cm000360;	gtagtctgcg	cm000358;
accctagttc	cm000360;	cctaagtacg	cm000360;	gactagccta	cm000360;	gtatagacct	cm000360;
acctaatacg	cm000360;	cctaagtcag	cm000360;	gactaggctg	cm000360;	gtatcccct	cm000360;

Продолжение таблицы 1.

Слово	Файлы	Слово	Файлы	Слово	Файлы	Слово	Файлы
acctagtgsa	cm000360;	cctaagttag	cm000360;	gactaggtcc	cm000359;	gtatgatccc	cm000360;
acctgcatag	cm000360;	cctactaaa	cm000360;	gactagtctt	cm000360;	gtcagagggg	cm000360;
accttagtcg	cm000360;	cctagactcc	cm000360;	gactatagcg	cm000360;	gtccacgtct	cm000360;
accttatgcc	cm000360;	cctaggacc	cm000356;	gactatagtc	cm000360;	gtcccctagg	cm000360;
acgagcatag	cm000360;	cctaggcact	cm000360;	gactatgctc	cm000360;	gtcccctga	cm000360;
acgagttacc	cm000360;	cctaggcctt	cm000360;	gactatgggg	cm000360;	gtccgtaggc	cm000360;
acgcataggg	cm000360;	cctaggctaa	cm000360;	gactccgctt	cm000360;	gtccgtttac	cm000360;
acgcctataa	cm000360;	cctagggaca	cm000360;	gactcctagg	cm000360;	gtctatagc	cm000360;
acgctagtca	cm000360;	cctaggggat	cm000360;	gactcctatg	cm000356;	gtcgcgactt	cm000360;
acggggcata	cm000360;	cctaggggcc	cm000358;	gactctacgt	cm000360;	gtcggagacc	cm000358;
acgtctatgc	cm000360;	cctaggggct	cm000358; cm000359;	gactgagtc	cm000357;	gtctaagacc	cm000360;
actaaggctc	cm000360;	cctaggtata	cm000360;	gactggtctc	cm000360;	gtctaggact	cm000360;
actaggcaag	cm000360;	cctaggtccc	cm000358;	gactggttac	cm000360;	gtctccctag	cm000360;
actaggcccc	cm000360;	cctagtacc	cm000360;	gacttagcct	cm000360;	gtctgaccat	cm000360;
actaggcctg	cm000356;	cctagtctta	cm000360;	gacttagggt	cm000360;	gtctgagtag	cm000360;
actagggata	cm000360;	cctagttcaa	cm000360;	gactgttagg	cm000360;	gtctgggcta	cm000360;
actagggtct	cm000360;	cctatagacc	cm000360;	gagaccccc	cm000359;	gtcttatgac	cm000360;
actataagcc	cm000360;	cctataggag	cm000360;	gagacctacc	cm000360;	gtgacctagt	cm000360;
actatagtag	cm000360;	cctatagggc	cm000360;	gagactcagt	cm000360;	gtgactagac	cm000360;
actatgggtc	cm000360;	cctatagagt	cm000360;	gagcctaggt	cm000360;	gtgactcgcc	cm000360;
actcgggggt	cm000360;	cctatgatcc	cm000360;	gagcctagtc	cm000360;	gtgagcctaa	cm000360;
actctagtcc	cm000360;	cctatgcggg	cm000360;	gaggactagt	cm000360;	gtgagtccta	cm000360;
actctctag	cm000360;	cctatgggcc	cm000360;	gagggtctaa	cm000360;	gtgataagtc	cm000360;
agacctacg	cm000360;	cctatgtgga	cm000360;	gagtagggcc	cm000359;	gtgatatccc	cm000360;
agacctgac	cm000358;	cctcagggcc	cm000360;	gagtcaccta	cm000360;	gtgccacgta	cm000360;
agacctgta	cm000360;	cctcataggg	cm000360;	gagtcgcat	cm000360;	gtgccctata	cm000360;
agacctgtc	cm000360;	cctcataggt	cm000360;	gagtgcccc	cm000360;	gtggcatagc	cm000360;
agacctgct	cm000360;	cctctagacc	cm000360;	gagttacccc	cm000360;	gttaggactg	cm000360;
agaccttata	cm000360;	cctctcagat	cm000360;	gatagcgtc	cm000360;	taaataagtc	cm000360;
agacggtctc	cm000360;	cctctgacaa	cm000360;	gatgtcctta	cm000360;	taaccgaga	cm000360;
agacgtctac	cm000360;	cctcttaagt	cm000360;	gattccctaa	cm000360;	taaccgggtg	cm000360;
agactaagtc	cm000358;	cctgacctct	cm000360;	gcaactaag	cm000360;	taacctagg	cm000360;
agactaggtc	cm000356;	ccttactagg	cm000360;	gccatagtc	cm000360;	taacctegg	cm000360;
agactagtgc	cm000360;	ccttagactg	cm000360;	gcccatagat	cm000360;	taacgtagac	cm000360;
agactatgct	cm000360;	ccttaggact	cm000360;	gccctaggt	cm000360;	taactaggcc	cm000360;
agactcacgg	cm000358;	cctgtagg	cm000360;	gccctacgt	cm000360;	taactcacgg	cm000360;
agactccgct	cm000360;	cgacataggg	cm000360;	gccctagga	cm000360;	taagagtcgg	cm000360;

Продолжение таблицы 1.

Слово	Файлы	Слово	Файлы	Слово	Файлы	Слово	Файлы
agagacctac	cm000360;	cgacccttag	cm000360;	gcccctgact	cm000360;	taaggcctc	cm000360;
agagacttag	cm000360;	cgacctaggt	cm000360;	gccctcatgg	cm000360;	taagtcagac	cm000360;
agagccccta	cm000360;	cgacgcccta	cm000360;	gcccttagtt	cm000360;	taattcgggg	cm000360;
agaggccccgc	cm000360;	cgacgtaatc	cm000360;	gccgactata	cm000360;	tacatagggc	cm000360;
agcccctaag	cm000360;	cgactaagtc	cm000359;	gcctaaggac	cm000360;	taccaagtcc	cm000360;
agcccctagg	cm000360;	cgactcgtat	cm000360;	gcctatactc	cm000360;	taccagtcg	cm000360;
agccctatta	cm000360;	cgagactatg	cm000360;	gcctataggg	cm000358;	taccccgtag	cm000360;
agccgtcaga	cm000360;	cgaggtacct	cm000360;	gcctataggt	cm000360;	tacctaaga	cm000360;
agcctatacg	cm000360;	cgaggtaac	cm000360;	gcctatgggc	cm000360;	tacctaggg	cm000358;
agcctccta	cm000360;	cgagtctcag	cm000360;	gcgaagtcta	cm000358;	tacctaagtc	cm000360;
agctaggact	cm000360;	cgagttacca	cm000360;	gcgacctaac	cm000356;	tacctcagtc	cm000360;
agctaggcca	cm000360;	cgatctaagg	cm000360;	gcgacctatg	cm000360;	taccttagac	cm000360;
aggagacctta	cm000360;	cgccctaggc	cm000358;	gcgacctatg	cm000360;	tacgaaagtc	cm000360;
aggagccctc	cm000360;	cgccctatag	cm000360;	gcgacctatgg	cm000360;	tacggaccct	cm000360;
aggagtcccc	cm000360;	cgccctatgt	cm000360;	gcgctagacc	cm000360;	tacggcgtag	cm000360;
aggccctagt	cm000358;	cgcgaggcct	cm000360;	gcgctctgag	cm000360;	tacgggtcac	cm000360;
aggccctcta	cm000358;	cgctaagtcg	cm000360;	gcgggtaata	cm000360;	tactgagaga	cm000360;
aggcctaagc	cm000360;	cggactatag	cm000360;	gcggtaacct	cm000360;	tacttaggag	cm000360;
aggctagtct	cm000360;	cggactatat	cm000360;	gcgtcatagc	cm000360;	tagacaggtc	cm000360;
agggacctat	cm000360;	cggagtatag	cm000360;	gcgtctaacc	cm000360;	tagaccgca	cm000360;
aggggcctat	cm000356;	cggataagag	cm000360;	gcgtctatag	cm000360;	tagacctcac	cm000360;
aggggactag	cm000360;	cgggactatg	cm000360;	gctacgtagt	cm000360;	tagactcgac	cm000360;
aggggctccgc	cm000358;	cgggactcag	cm000360;	gctatagacc	cm000360;	tagactcgag	cm000360;
aggggctgccc	cm000360;	cgggcctcat	cm000360;	gctatagggc	cm000360;	tagacttacc	cm000360;
agggtacctc	cm000359;	cggggctaca	cm000360;	gctcgtagac	cm000360;	tagagtaacc	cm000360;
agggtatgcc	cm000360;	cgggggattc	cm000360;	gcttagactg	cm000360;	tagagtactc	cm000360;
agggtatgtc	cm000360;	cggggtacta	cm000360;	ggaacctaa	cm000360;	tagatctacg	cm000360;
agggtcctga	cm000360;	cggggtatac	cm000360;	ggacccccca	cm000360;	tagccccctg	cm000360;
aggtagggtc	cm000360;	cgggtaccta	cm000358;	ggacctatg	cm000356;	tagcctaggg	cm000360;
aggccctac	cm000360;	cgggctcta	cm000360;	ggacgctagt	cm000360;	tagcctcctg	cm000360;
aggttactag	cm000360;	cgggttactc	cm000360;	ggactaagac	cm000360;	tagcctcgac	cm000359;
agtaccccg	cm000360;	cggtctcaga	cm000360;	ggactagccc	cm000358;	taggactagc	cm000360;
agtacctagc	cm000358;	cggtcttagg	cm000360;	ggactagtcc	cm000360;	taggagtcg	cm000358;
agtacacctc	cm000360;	cgtaaggcct	cm000360;	ggactcagtg	cm000360;	taggccccag	cm000360;
agttagactta	cm000360;	cgtaagtcc	cm000360;	ggactccacg	cm000360;	taggcccagt	cm000358;
agttagccta	cm000358;	cgtagacctta	cm000358;	ggactcctgt	cm000360;	taggcgaggt	cm000360;
agtcacgagg	cm000358;	cgtagacctc	cm000360;	ggacttagtc	cm000360;	taggcgatac	cm000360;
agtcatacct	cm000360;	cgtaggctac	cm000360;	ggacttctag	cm000360;	taggcggagc	cm000358;

Продолжение таблицы 1.

Слово	Файлы	Слово	Файлы	Слово	Файлы	Слово	Файлы
agtccctaag	cm000357;	cgtaagacc	cm000360;	ggagactcta	cm000360;	taggctagtc	cm000360;
agtcttaggc	cm000358;	cgtagactaca	cm000360;	ggagtcacct	cm000360;	taggctact	cm000360;
agtgacctac	cm000360;	cgtagggggcc	cm000360;	ggagtccta	cm000358;	tagggacct	cm000360;
agttacccc	cm000360;	cgtaggacct	cm000360;	ggacttagt	cm000360;	tagggacctg	cm000360;
ataagtcgc	cm000360;	ctaagacc	cm000360;	ggacctaga	cm000360;	tagggctata	cm000360;
ataccctaag	cm000360;	ctaaggaggc	cm000360;	ggccatgtga	cm000360;	taggggcta	cm000356;
ataccctagt	cm000360;	ctaaggcggg	cm000360;	ggcccgttag	cm000360;	taggggctct	cm000360;
ataccctgcc	cm000360;	ctaaggcgtc	cm000358;	ggccatagg	cm000360;	taggggcta	cm000360;
atacgacct	cm000360;	ctaagggtcg	cm000360;	ggccatac	cm000360;	taggggtcga	cm000359;
atacgcgta	cm000360;	ctaagtacgg	cm000360;	ggcctctacg	cm000360;	tagggtcta	cm000360;
atagaccgc	cm000360;	ctaagtccga	cm000360;	ggcctctgag	cm000360;	taggtaagtc	cm000360;
atagaccctc	cm000360;	ctacctaggc	cm000360;	ggccttagga	cm000358;	taggtacccc	cm000360;
atagacgggt	cm000360;	ctacggaggt	cm000360;	ggcctctga	cm000360;	taggtactca	cm000360;
atagaggatc	cm000360;	ctagaccgc	cm000360;	gggatccta	cm000360;	taggtccca	cm000358;
atagcctgcc	cm000360;	ctagacctcc	cm000360;	ggctaagcac	cm000360;	taggtccccc	cm000358;
ataggagtcc	cm000360;	ctagagtccg	cm000356;	ggctacgtag	cm000360;	taggtccctg	cm000358;
ataggccctc	cm000360;	ctagataggc	cm000360;	ggctatacct	cm000360;	taggtgagtc	cm000360;
ataggggcct	cm000359;	ctaggcaagg	cm000360;	ggctatagcc	cm000360;	tagtcagggc	cm000360;
atagggctc	cm000360;	ctaggcccgt	cm000360;	ggctatagg	cm000360;	tagtccagg	cm000360;
ataggtcgta	cm000360;	ctaggacc	cm000356;	ggctataag	cm000360;	tagtctgtc	cm000360;
atatggggag	cm000360;	ctagggtcc	cm000358; cm000360;	gggacctag	cm000360;	tagtctgta	cm000360;
atcaggcctc	cm000360;	ctagggtact	cm000356;	gggacctagg	cm000360;	tataactgcg	cm000360;
atcatagggg	cm000360;	ctagggtccg	cm000358;	gggactagtc	cm000360;	tatacgccgt	cm000360;
atcgaggta	cm000360;	ctaggctcct	cm000358;	gggacttcta	cm000360;	tatacggtc	cm000360;
atctcggtc	cm000360;	ctaggctggg	cm000358;	gggagtccta	cm000356;	tatactgtc	cm000360;
atcttagggg	cm000360;	ctagtatacg	cm000360;	gggcataagt	cm000360;	tatactggca	cm000360;
atgagtagac	cm000360;	ctagtcagg	cm000360;	gggcatagg	cm000360;	tatagacc	cm000360;
atgcgtagtc	cm000360;	ctagtctgc	cm000360;	gggcctcta	cm000360;	tataggactc	cm000360;
atggggccta	cm000360;	ctagtctcg	cm000360;	gggcctatag	cm000358;	tatagggtcc	cm000360;
caacttagtg	cm000360;	ctatagacc	cm000360;	gggcctcaga	cm000360;	tatagtcgg	cm000360;
caagatccgt	cm000360;	ctatagggcg	cm000360;	gggcgctaa	cm000357;	tatagggtc	cm000360;
caaggcctat	cm000360;	ctataggtc	cm000360;	gggcgtagcc	cm000360;	tcacctagg	cm000360;
cagacctga	cm000360;	ctatagtcgg	cm000360;	gggctatag	cm000360;	tcacgggett	cm000360;
cagagtcta	cm000359;	ctatggacc	cm000360;	gggcttagac	cm000360;	tcacgtgact	cm000360;
caggaccca	cm000360;	ctatgggagt	cm000360;	ggggagtcta	cm000358;	tcagaccga	cm000360;
caggggggtc	cm000360;	ctcagaggcc	cm000358;	ggggcatgac	cm000360;	tcagagacta	cm000360;
caggtacccc	cm000360;	ctcagctagg	cm000360;	gggggtccta	cm000360;	tcagggctc	cm000360;

Продолжение таблицы 1.

Слово	Файлы	Слово	Файлы	Слово	Файлы	Слово	Файлы
cagctctagg	cm000358;	ctcaggagtc	cm000360;	ggggtccgta	cm000360;	tcaggtctgc	cm000358;
cataaggacc	cm000360;	ctcagggggc	cm000360;	ggggtctagt	cm000360;	tccagaggtta	cm000360;
cataggccta	cm000358;	ctcagtgggc	cm000360;	ggggttaacc	cm000360;	tcctaagtc	cm000360;
catagggcac	cm000360;	ctcatagagt	cm000360;	gggtactctg	cm000360;	tcctatagc	cm000360;
catagggggt	cm000360;	ctcatgggcc	cm000360;	gggtatcccc	cm000360;	tccgactagg	cm000358;
cataggtata	cm000360;	ctccctaggt	cm000358;	gggtatccta	cm000360;	tccgctata	cm000360;
cataggtatc	cm000360;	ctccgtaggg	cm000360;	gggtcacgta	cm000360;	tcctagggga	cm000360;
cataggtcgt	cm000360;	ctcctaggcc	cm000360;	gggtcctgac	cm000360;	ttagaccgcg	cm000360;
catataggag	cm000360;	ctcctagtgg	cm000360;	gggtccttat	cm000358;	ttaggggtgg	cm000360;
catggcgtct	cm000360;	ctccttaggc	cm000359;	gggtctaacg	cm000360;	ttcacctag	cm000360;
catgtctagg	cm000360;	ctcgtataag	cm000360;	gggtctcact	cm000360;	tctcatagac	cm000360;
ccaaccttag	cm000360;	ctctacgtat	cm000360;	gggttactac	cm000360;	tctcgtata	cm000360;
ccaagccta	cm000360;	ctctagtcga	cm000360;	ggtacctaga	cm000360;	tctgagacc	cm000358;
ccactaggct	cm000360;	ctctgacaag	cm000360;	ggtactcagg	cm000360;	tctgataccc	cm000360;
ccagacttaa	cm000360;	ctctgagact	cm000360;	ggtacttagg	cm000360;	tctgtggact	cm000360;
ccagagactt	cm000360;	ctgagacccc	cm000356;	ggtagacccc	cm000360;	tgacatagg	cm000356;
ccaggtctcc	cm000360;	ctgaggggtc	cm000360;	ggtagtatgc	cm000360;	tgaccctca	cm000360;
ccatagacgg	cm000360;	ctgagtccgc	cm000360;	ggtagacccc	cm000359;	tgagactccg	cm000360;
ccatagggga	cm000360;	ctgggcctag	cm000358;	ggtagacggg	cm000360;	tgaggaccgc	cm000356;
cccataggct	cm000360;	ctggggatc	cm000360;	ggtagagccc	cm000360;	tgaggcgtta	cm000360;
cccacgtag	cm000360;	ctggggctta	cm000360;	ggtagacccc	cm000360;	tgagggacta	cm000360;
cccctaagg	cm000360;	ctgggtacct	cm000360;	ggtagactta	cm000360;	tgagtaccgc	cm000358;
cccctaagt	cm000360;	cttagacgga	cm000360;	ggtagtactc	cm000360;	tgagtccgcc	cm000360;
cccctaagg	cm000360;	cttagcccct	cm000360;	ggtagtaggg	cm000360;	tgatccgtag	cm000360;
cccctaaggt	cm000360;	cttagggaac	cm000360;	ggtagatagc	cm000360;	tgagacccta	cm000360;
cccctaggct	cm000360;	cttagggacc	cm000360;	ggtagatacc	cm000358;	tgaggcgtgac	cm000360;
cccctaggtc	cm000360;	cttaggggac	cm000360;	ggtagcctact	cm000360;	tgaggctacta	cm000360;
cccctctatg	cm000360;	cttaggggct	cm000358;	ggtagctgagt	cm000360;	tgaggctaac	cm000360;
cccgacttag	cm000358;	cttatgaggc	cm000360;	ggtagtaacc	cm000360;	tggtctccta	cm000360;
cccgacttat	cm000358;	gaaccttagg	cm000360;	ggtagtagcat	cm000360;	ttaacctggg	cm000360;
cccgatacag	cm000360;	gaaccttagt	cm000358;	ggtagtcaatg	cm000360;	ttacgcgtct	cm000360;
cccgctaaag	cm000360;	gaagaccccc	cm000360;	ggtagtcagcc	cm000360;	ttacggccga	cm000360;
ccctaatagt	cm000360;	gaataaggcc	cm000360;	ggtagtccgta	cm000360;	ttacggcgta	cm000360;
ccctacgagt	cm000360;	gacagaccac	cm000360;	ggtagtctagt	cm000360;	ttagggacc	cm000360;
ccctacgtcg	cm000360;	gaccactagg	cm000360;	ggtagctctata	cm000360;	ttagtctagg	cm000360;
ccctagacta	cm000360;	gacccccag	cm000360;	ggtagtgaagg	cm000360;	ttgatcccg	cm000360;
ccctagggtc	cm000358;	gaccctagat	cm000358;	ggtagttagga	cm000360;	ttgcgactca	cm000360;
ccctagtcca	cm000360;	gaccctccat	cm000360;	ggtagtccctga	cm000360;		

Окончание таблицы 1.

Слово	Файлы	Слово	Файлы	Слово	Файлы	Слово	Файлы
ccctatagcg	cm000360;	gaccctcgta	cm000360;	gtaacccggt	cm000360;		

Полный список потерянных слов минимальной длины вида *Arabidopsis thaliana* приведен в формате стека с файлами в таблицах 2 и 3.

Таблица 2 – *Arabidopsis thaliana*, длина потерянных слов – 8.

Слово	Файлы	Слово	Файлы	Слово	Файлы	Слово	Файлы
Aagcgcg	AJ270058;	ccgcgggc	AJ270058;	gcgcagcg	AJ270058;	gggcccgc	AJ270058;
Accccgcg	AJ270058;	ccggccgc	AJ270058;	gcgccaga	AJ270058;	gggcccgt	AJ270058;
Acccgggg	AJ270058;	cctacgac	AJ270058;	gcgcccc	AJ270058;	gggcgcct	AJ270058;
Acgcgga	AJ270058;	cctggggc	AJ270058;	gcgccccg	AJ270058;	gggcgcgc	AJ270058;
Acgtcg	AJ270058;	cgcacgga	AJ270058;	gcgcccgc	AJ270058;	gggcggtta	AJ270058;
Agcccccg	AJ270058;	cgcccc	AJ270058;	gcgcccgg	AE005173;	ggggcccg	AJ270058;
Agcgcg	AJ270058;	cgccccgt	AJ270058;	gcgcgacc	AJ270058;	ggggccgc	AJ270058;
agcgggg	AJ270058;	cgcccctg	AJ270058;	gcgcgcc	AJ270060;	ggggcgcc	AJ270058;
Aggcccc	AJ270058;	cgcccggc	AJ270058;	gcgcgcgg	AJ270058;	ggggcccg	AJ270058;
Aggcccgc	AJ270058;	cgcccc	AJ270058;	gcgcggcg	AJ270058;	gggggcgc	AJ270058;
Agggccgt	AJ270058;	cgcgcat	AJ270058;	gcgcgggt	AJ270058;	gggggctc	AJ270058;
agggcg	AJ270058;	cgcgggg	AJ270058;	gcgcgtag	AJ270058;	ggggggcc	AJ270058;
aggggcc	AJ270058;	cgcgctgg	AJ270058;	gcgcgtcc	AJ270058;	gggggggc	AJ270058;
Agggggtta	AJ270058;	cgccgta	AJ270058;	gcgccccg	AJ270058;	ggtgcgca	AJ270058;
Atacgg	AJ270058;	cgcgccc	AJ270058;	gcgggcc	AJ270058;	gtgccccg	AJ270058;
Atcgccg	AJ270058;	cgggacgc	AJ270058;	gcggcggt	AJ270058;	gtgcgcac	AJ270058;
cagcg	AJ270058;	cgggccc	AJ270058;	gcggggg	AJ270058;	gttacgcg	AJ270058;
Cagggccc	AJ270058;	cggggcgc	AJ270058;	gcggtgcc	AJ270058;	tacgcggg	AJ270058;
cagggcg	AJ270058;	cgggggc	AJ270058;	gcgtcg	AJ270058;	tagcgggc	AJ270058;
Ccccccg	AJ270058;	cgggggtc	AJ270058;	gcgtgcac	AJ270058;	taggcccc	AJ270058;
Cccccgg	AJ270058;	cgtaggg	AJ270058;	gcgtggcg	AJ270058;	tattgcgc	AJ270058;
Cccccgac	AJ270058;	cggtcg	AJ270058;	ggcccc	AE005173;	tccgcgcg	AJ270058;
Ccccggt	AJ270058;	ctgcggcc	AJ270058;	ggccccgc	AJ270058;	tcgcgcgc	AJ270058;
Cccgggg	AE005172;	ctgggcgc	AJ270058;	ggcccggt	AJ270058;	tcgtgcgc	AJ270058;
Cccgccc	AJ270058;	cttagcgc	AJ270058;	ggccgctc	AJ270058;	tgcccggg	AJ270058;
Cccgcg	AJ270058;	gacgcacg	AJ270058;	ggcgtgc	AJ270058;	tgcgggc	AJ270058;
Cccggcg	AJ270058;	gcacgccc	AJ270058;	ggcggccc	AJ270058;	tgcgggg	AJ270058;
Ccctgg	AJ270058;	gccccag	AJ270058;	ggcgccct	AJ270058;	tgcgctc	AJ270058;

Окочание таблицы 2

Слово	Файлы	Слово	Файлы	Слово	Файлы	Слово	Файлы
Ссctgcgc	AJ270058;	gcccgcgc	AJ270058;	ggcgcgcg	AJ270058;	ttcgcgcg	AJ270058;
Ссctgggc	AJ270058;	gcccgcgt	AJ270058;	ggcgcgct	AJ270058;	ttgcgcgt	AJ270058;
Ссgcсcta	AJ270058;	gссgcсct	AJ270058;	gggcccc	AJ270058;		
Ссgcgcgc	AJ270058;	gcctgggg	AJ270058;	gggccccg	AJ270058;		

Таблица 3–*Arabidopsis thaliana*, длина потерянных слов – 9.

Слово	Файлы	Слово	Файлы	Слово	Файлы	Слово	Файлы
aacggggcg	BA000015;	cctggggcg	BA000015;	gccccgct	BA000014;	ggcgggcag	BA000014;
acagcgggc	BA000014;	cgaccccc	BA000015;	gccccggc	BA000015;	ggcgtgcgc	BA000015;
acagggcgc	BA000014;	cgagcgcgc	BA000014;	gccccgcgc	BA000014;	gggccccg	BA000014;
accccccg	BA000014;	cgaggcccc	BA000014;	gcccgcct	BA000014;	gggccccga	BA000015;
acccccgg	BA000015;	cgcccccg	BA000015;	gcccgcgc	BA000014;	gggcccct	BA000014;
accgcgccc	BA000015;	cgccccgat	BA000015;	gcccgcgcg	BA000015;	gggcccgc	BA000014;
accggcccc	BA000015;	cgccccggt	BA000015;	gcccgcgct	BA000014;	gggcccggg	BA000014;
acgccccgt	BA000014;	cgccccgcg	BA000015;	gcccggcgt	BA000014;	gggcccgcg	BA000014;
acgcccggc	BA000014;	cgcccgcgt	BA000014;	gcccgggc	BA000015;	gggcgcccg	BA000015;
acgcccgggt	BA000015;	cgcccgggg	BA000015;	gcccgtgc	BA000015;	gggcgcgaa	BA000015;
acgcgcgcc	BA000014;	cgccctacg	BA000015;	gcccggcg	BA000014;	gggcgcgca	BA000015;
acgcgcgct	BA000015;	cgcgacct	BA000015;	gcccgcggg	BA000014;	gggcgcgcc	BA000014;
acgcggggg	BA000014;	cgcgcacgc	BA000014;	gcccgcgtg	BA000014;	gggcgcgcg	BA000015;
acgggcccc	BA000015;	cgcgcacgg	BA000015;	gcccggcgc	BA000015;	gggcgggcg	BA000014;
acggggccc	BA000015;	cgcgccccg	BA000014;	gcctggccc	BA000015;	gggcgtccg	BA000015;
acggggcgc	BA000014;	cgcgcccgg	BA000014;	gcgacgcgc	BA000015;	ggggcccc	BA000015;
acggggcgt	BA000015;	cgcgcccgt	BA000015;	gcgaccgc	BA000015;	ggggccccg	BA000014;
acggggggc	BA000014;	cgcgcgacg	BA000015;	gcgcccca	BA000014;	ggggcccgc	BA000015;
acggggggg	BA000015;	cgcgcgccc	BA000014;	gcgccccg	BA000015;	ggggcccgg	BA000015;
agccccccc	BA000015;	cgcgcgсga	BA000015;	gcgccccga	BA000014;	ggggcccgt	BA000015;
agccccgcg	BA000014;	cgcggggac	BA000014;	gcgccccgc	BA000014;	ggggcccgcg	BA000015;
agcggcccc	BA000014;	cgcgggggc	BA000015;	gcgccccgg	BA000015;	ggggcccgtc	BA000015;
agcggccccg	BA000014;	cgcgggggg	BA000015;	gcgccccgt	BA000015;	ggggcgcac	BA000014;
agcggccccg	BA000014;	cgcggggta	BA000014;	gcgccccgtg	BA000015;	ggggcgccc	BA000015;

Продолжение таблицы 3.

Слово	Файлы	Слово	Файлы	Слово	Файлы	Слово	Файлы
agcgcggac	BA000015;	cgcgctccg	BA000014;	gcgcccgga	BA000015;	ggggcgcgc	BA000015;
agcgcgggg	BA000015;	cgcgctgac	BA000014;	gcgcccggc	BA000015;	ggggcgcgg	BA000014;
agcggggcg	BA000014;	cgcggccc	BA000014;	gcgcccggg	BA000014;	gggggcccc	BA000015;
aggcccc	BA000015;	cgcggcctg	BA000015;	gcgcccgg	BA000014;	gggggccc	BA000015;
aggccccg	BA000014;	cgcggggca	BA000015;	gcgcccgt	BA000015;	gggggcccga	BA000014;
aggccccg	BA000014;	cgcggggcg	BA000015;	gcgcccgt	BA000015;	gggggccc	BA000015;
aggggggg	BA000015;	cgctgcgcg	BA000015;	gcgcccctg	BA000015;	gggggcgcg	BA000014;
aggggccc	BA000015;	cggacgggg	BA000015;	gcgcgacag	BA000015;	gggggcgct	BA000014;
aggggcgcc	BA000015;	cggagccc	BA000015;	gcgcgcacg	BA000015;	ggggggccc	BA000015;
aggggggcc	BA000015;	cggcccc	BA000015;	gcgcgact	BA000015;	ggggggccg	BA000015;
aggggggcg	BA000014;	cggccccg	BA000015;	gcgcgccc	BA000014;	ggggggcgc	BA000015;
aggggtcgc	BA000014;	cggccccg	BA000015;	gcgcgccc	BA000014;	ggggggcgg	BA000014;
agtgcgcgc	BA000015;	cggcccgcg	BA000014;	gcgcgccc	BA000014;	gggggggccc	BA000015;
ataggcgc	BA000015;	cggccgcgc	BA000015;	gcgcgccc	BA000015;	gggggggccc	BA000015;
atagtgcgc	BA000015;	cgggcccga	BA000015;	gcgcgcgac	BA000015;	ggggtgcgc	BA000015;
atccgtgcg	BA000015;	cgggccc	BA000014;	gcgcgcggc	BA000015;	gggtagcgc	BA000014;
atcgcccc	BA000014;	cgggcccct	BA000015;	gcgcgcgta	BA000015;	gggtccc	BA000015;
atcgggggg	BA000015;	cgggcccgc	BA000015;	gcgcgctac	BA000015;	gggtccc	BA000014;
atcgcggc	BA000015;	cgggcccgt	BA000014;	gcgcgctcc	BA000014;	gtaccccc	BA000015;
atggggcg	BA000014;	cgggcccgc	BA000014;	gcgcgctcg	BA000015;	gtaggccc	BA000015;
cacgcgggc	BA000014;	cgggcccct	BA000015;	gcgcggaca	BA000015;	gtaggcgc	BA000014;
cacggcg	BA000014;	cgggcccga	BA000015;	gcgcggcct	BA000015;	gtcccc	BA000014;
caggcccgcg	BA000015;	cgggcccgc	BA000015;	gcgcgggac	BA000014;	gtcccgcg	BA000014;
caggccgtc	BA000014;	cgggcccgc	BA000014;	gcgcggggc	BA000015;	gtccgggg	BA000015;
cagtgcgcg	BA000014;	cggggccc	BA000015;	gcgcggggt	BA000015;	gtcgcgcc	BA000015;
catagcgcg	BA000015;	cggggccc	BA000015;	gcgcgggtc	BA000015;	gtcgggg	BA000014;
cccaggggc	BA000015;	cggggccc	BA000015;	gcgctacc	BA000014;	gtcggggg	BA000015;
ccccccag	BA000014;	cggggcgac	BA000015;	gcgctagcg	BA000014;	gtgccccc	BA000015;
ccccccgt	BA000015;	cggggcgca	BA000014;	gcgctccgc	BA000015;	gtgcccc	BA000014;

Продолжение таблицы 3.

Слово	Файлы	Слово	Файлы	Слово	Файлы	Слово	Файлы
cccccggg	BA000014; BA000015;	cggggcgct	BA000014;	gcggaccgc	BA000014;	gtgcgcgct	BA000015;
cccccgta	BA000015;	cgggggccc	BA000015;	gcggcccgc	BA000014;	gtcggggcg	BA000015;
cccccgtc	BA000015;	cgggggcct	BA000015;	gcggcctag	BA000015;	gtcggggcg	BA000015;
ccccggcc	BA000014;	cgggggcgc	BA000014; BA000015;	gcgggcccc	BA000015;	taccccccg	BA000015;
ccccgggc	BA000014; BA000015;	cgggggcg	BA000014;	gcgggcccg	BA000015;	taccgcgcg	BA000014;
ccccgctc	BA000015;	cgggggggc	BA000014; BA000015;	gcgggcgcc	BA000014;	tacgggcgc	BA000014;
ccccgcta	BA000015;	cgggggggt	BA000014;	gcgggcgcg	BA000015;	tagccccc	BA000014;
ccccctacg	BA000014;	cggggggta	BA000014;	gcggggccc	BA000015;	taggcacgc	BA000014;
ccccgcgcg	BA000014;	cggggggtg	BA000014;	gcggggcct	BA000015;	tagggcgcc	BA000015;
ccccggggg	BA000015;	cggtcccc	BA000014;	gcgggggcc	BA000015;	taggggcgc	BA000015;
ccccgggtg	BA000014;	cgccgggc	BA000014;	gcggggtgc	BA000014;	tagtgcgcg	BA000015;
ccccgtcgc	BA000014;	cgccgcgcc	BA000015;	gcgggtcgc	BA000015;	tcccgcgcg	BA000014;
cccgcgcgc	BA000014; BA000015;	cgccgcgcc	BA000015;	gcgtaccgc	BA000015;	tccgcgcgc	BA000015;
cccgcgcgg	BA000015;	cgccgcgcg	BA000015;	gcgtcccgc	BA000015;	tccgcgcgg	BA000014;
cccgcgcgt	BA000014;	cgccgcgtac	BA000014;	gcgtgcgca	BA000014;	tccgtgcgg	BA000015;
cccgtcggg	BA000015;	ctgcccccc	BA000015;	ggaccccgc	BA000015;	tcgcccccc	BA000015;
cccgggcgc	BA000015;	ctgcccccc	BA000014; BA000015;	ggacgggcg	BA000015;	tcgcccccc	BA000015;
cccgggcgg	BA000015;	ctgcccgcg	BA000015;	ggagcgcac	BA000014;	tcgcccgcg	BA000014;
ccctgcgcg	BA000014;	ctgcccgcg	BA000014; BA000015;	ggccccccc	BA000015;	tcgcgctcg	BA000015;
ccgcgcccg	BA000015;	ctggcccccc	BA000014;	ggccccccg	BA000014;	tcgggcccc	BA000014;
ccgcgcgga	BA000014;	ctggcgccc	BA000014;	ggccccccg	BA000014; BA000015;	tgccccgcg	BA000015;
ccgcgcggc	BA000015;	ctgggcggg	BA000014;	ggcccgcag	BA000015;	tgccccgcg	BA000015;
ccgggcgca	BA000015;	ctggggcg	BA000014;	ggcccgcgc	BA000014;	tgcccct	BA000015;
ccgggcgcc	BA000014; BA000015;	gacgggggg	BA000015;	ggcccggag	BA000014;	tgcccggg	BA000014; BA000015;
ccgggcgcg	BA000014; BA000015;	gagcgcgca	BA000014;	ggcccggca	BA000014;	tgcccgcg	BA000015;
ccgggcgtg	BA000014;	gagcggggc	BA000015;	ggcccgcg	BA000015;	tgcccgcg	BA000015;
ccggggcgc	BA000014;	gcaccgcgc	BA000014;	ggcgcacgg	BA000015;	tgcccgcgt	BA000015;

Окочание таблицы 2

Слово	Файлы	Слово	Файлы	Слово	Файлы	Слово	Файлы
ccgggggccc	BA000015;	gcacgcgac	BA000015;	ggcgccccg	BA000014; BA000015;	tgcgcgggg	BA000015;
ccggggggg	BA000015;	gcacgcggg	BA000015;	ggcgcccga	BA000015;	tgcgggccc	BA000014;
ccgggggggt	BA000014;	gcaggcgc	BA000015;	ggcgcgca	BA000015;	tgcgggccc	BA000014;
ccgtcgccc	BA000015;	gccccagc	BA000014;	ggcgcgctc	BA000015;	tgcgggccc	BA000015;
ccgtcgcg	BA000014;	gccccccg	BA000015;	ggcgcgcc	BA000015;	tgggcgcg	BA000015;
cctcgggcc	BA000015;	gccccccg	BA000015;	ggcgcggt	BA000014;		
cctcgcg	BA000015;	gccccccgt	BA000014; BA000015;	ggcgcgctc	BA000014;		
cctcgcgcc	BA000014;	gccccccgat	BA000014;	ggcgggccc	BA000014;		

Полный список потерянных слов минимальной длины некоторых хлоропластов приведен в формате стека с файлами в таблицах 4, 5 и 6. Ввиду большого количества файлов, в таблицах 5 и 6 будет отражаться их количество.

Таблица 4 – хлоропласты, длина потерянных слов – 5.

Слово	Файлы
cgccc	U87145;
ggccc	U87145;

Таблица 5 – хлоропласты, длина потерянных слов – 6.

Слово	n	Слово	n	Слово	n	Слово	n	Слово	n	Слово	n	Слово	n
aaaaat	1	ccccga	2	cgcaaga	1	cgtacg	1	gcccct	1	gcgtgc	1	gggtcc	1
aaggcg	1	ccccgc	2	cgcccg	1	cgtccc	1	gcccga	1	gcgtgg	1	gtccc	1
acagcg	1	ccccgg	3	cgccga	1	cgtcgc	1	gcccgg	10	gctagg	13	ggtggc	1
accccg	2	ccctc	14	cgccgg	1	cgtcgg	1	gcccgt	1	gctcgg	1	gtacc	1
acgccg	4	cccag	1	cgcgac	1	cgtgcc	2	gcccctc	3	gctgtg	1	gtccc	1
acgcga	1	cccgcg	3	cgcgag	2	cgtggc	2	gcccga	1	ggacgc	1	gtccga	1
acgcgc	2	cccgt	1	cgcgat	4	cgtggg	1	gcccgcg	2	ggagcc	1	gtccgc	1
acgcgg	1	cccggc	3	cgcgca	1	cgctcg	1	gcccga	1	ggagcg	1	gtccgg	2
acggac	1	cccggg	1	cgcgcc	1	ctaggg	1	gcccgc	13	ggcaca	1	gtcgcg	2
acggtg	2	cccgtg	3	cgcgcg	26	ctcgca	1	gcccggg	5	ggcacc	1	gtcggc	3
acgttg	1	ccctcc	1	cgcgct	1	ctcgga	1	gcccgt	1	ggccac	2	gtctcg	1
actcgc	2	ccctcg	2	cgcggc	1	ctgegg	1	gcccgtc	3	ggccca	1	gtgccc	2
agcccc	1	ccgacg	1	cgcggg	4	ctggcg	1	gcccctg	14	ggcccg	1	gtgccc	2

Окончание таблицы 5

Слово	n	Слово	n	Слово	n	Слово	n	Слово	n	Слово	n	Слово	n
agcccg	1	ccgagc	1	cgcggt	1	ctgtgg	1	gcgaag	1	ggccct	1	gtgcgc	7
agccgg	1	ccgata	1	cgcgtc	1	gacccc	2	gcgaga	1	ggccga	1	gtggcc	1
agcgcc	3	ccgcag	2	cgcggt	2	gaccgc	11	gcgagg	1	ggccgc	2	gtgtcg	1
agcgcg	2	ccgcgc	6	Cgctgc	1	gaccgg	1	gcgata	1	ggccgg	5	taaagt	1
agcggg	1	cccggt	2	Cgctgt	1	gacctc	1	gcgcac	9	ggcctc	1	tatttt	1
aggccg	1	ccgctc	1	cggacc	2	gacgcc	1	gcgcag	1	ggcctg	1	tcacgc	1
aggcgc	1	ccggac	2	cggacg	1	gacgcg	1	gcgcat	4	ggcgcc	14	tccttg	1
agggca	1	ccggcc	6	cggcag	1	gacgga	1	gcgcca	1	ggcgcg	2	tccgcg	2
aggggg	1	ccggcg	8	cggccc	3	gacggc	3	gcgccg	4	ggcgct	2	tcgagc	1
atggcg	1	ccgggc	1	cggccg	2	gagagc	1	gcgcga	1	ggcgtc	3	tcgagg	2
attaat	1	ccgggg	2	cggcgc	1	gagccc	1	gcgcgc	18	ggctcg	1	tcggac	1
cacacg	1	ccgtcc	1	cggcgt	8	gagcgg	1	gcgcgg	5	ggctct	1	tcggcc	1
cacgcg	1	ccgtgt	1	cgggag	1	gaggcc	2	gcgcgt	4	gggacc	1	tcgggg	1
cacggg	2	cctagc	13	cgggca	1	gaggcg	1	gcgctg	2	gggact	2	tctcgg	1
cagagc	1	cctcgg	2	cggggc	2	gagggc	1	gcggcc	2	gggagc	1	tgccgg	1
cagcgc	2	cgacgc	4	cggggg	1	gcacag	1	gcggcg	2	gggccg	5	tgcggc	1
caggcc	1	cgagcg	1	cgggggt	2	gcacgg	1	gcgggg	5	gggcgc	13	tggccc	1
catcgc	2	cgaggc	1	Cggtct	1	gccacc	1	gcggtc	5	ggggcg	12	tggcgc	1
ccacgg	1	cgagtg	1	Cggtga	1	gccacg	1	gcgtcc	1	gggggg	2	tgggcc	3
cccacg	1	cgatag	1	Cggtgc	1	gccatc	1	gcgtcg	1	ggggtc	7	tgcccg	1
ccccc	9	cgatcg	3	Cggtgg	1	gccccg	4	gcgtga	1	ggggtg	1		

Таблица 6 – хлоропласты, длина потерянных слов – 7.

Слово	n	Слово	n	Слово	n	Слово	n	Слово	n	Слово	n	Слово	n
aaacacg	1	aggcgcg	44	ccgatct	4	cgctcgt	34	gcagcgg	4	ggcgctcc	9	tagcgac	5
aaaccgg	7	aggcggc	59	ccgatgc	1	cgctcta	4	gcagcgt	50	ggcgctcg	19	tagcgag	13
aaacgcc	21	aggcggt	1	ccgatgg	3	cgctctg	4	gcagctc	5	ggcgctct	24	tagcgca	28
aaacgcg	1	aggctcc	2	ccgatgt	1	cgctcac	1	gcaggac	1	ggcggtga	6	tagcgcg	64
aaacggc	4	aggctgg	1	ccgcaac	1	cgctcga	22	gcaggag	1	ggcggtgc	2	tagcgct	69
aaacgtg	2	aggctgt	15	ccgcaga	11	cgctgcc	6	gcaggcc	3	ggcggtgg	4	tagcgga	7
aaacacg	1	aggggacc	1	ccgcagc	1	cgctcgcg	96	gcagggc	4	ggcggtgt	67	tagcgta	56
aaacacc	49	aggggac	66	ccgcatg	2	cgctcggc	26	gcagggtg	47	ggcggttg	20	tagcgtc	11
aaacaccg	8	aggggccc	1	ccgcatt	1	cgctcggg	18	gcagctcc	13	ggcggttt	6	tagcgtg	86
aaacacgc	47	aggggcga	5	ccgccaa	1	cgctcggg	54	gcagctcg	28	ggcggtacc	1	tagcggt	61
aaacacgt	10	aggggcgc	1	ccgcacg	1	cgctctaa	1	gcagctct	1	ggcggtagc	19	tagcgagc	42
aaacactc	1	aggggcgg	4	ccgcctc	3	cgctctac	47	gcagctgt	5	ggcggtcca	15	tagcgccg	57
aaacactg	45	aggggctg	12	ccgcgca	1	cgctctag	3	gcactact	2	ggcggtctg	2	tagcggg	1

Продолжение таблицы 6.

Слово	n	Слово	n	Слово	n	Слово	n	Слово	n	Слово	n	Слово	N
aacagcg	23	aggggac	1	ccgccgc	12	cgctccc	1	gcatagt	5	ggctgcc	86	taggcta	7
aacagtg	5	aggggtt	1	ccgccgg	15	cgctccg	5	gcatccg	1	ggctgcg	25	taggctc	1
aacatcg	4	aggggtac	2	ccgcctc	1	cgctcgc	12	gcatcgt	16	ggctggc	22	tagggac	36
aaccgcc	3	agggtag	4	ccgcctt	1	cgctcgg	9	gcatgac	1	ggctggg	4	tagggcc	5
aaccgcg	9	aggggtca	1	ccgcgac	43	cgctcta	6	gcatgcg	17	ggctggt	1	tagggcg	51
aaccggc	35	aggggtcc	1	ccgcgag	6	cgctctc	35	gcattag	1	ggctgtg	1	taggggg	35
aaccgtc	1	aggggtgc	1	ccgcgat	7	cgtgaca	1	gcatcgc	17	ggcttaa	24	taggtcg	2
aaccgtt	1	aggggtgt	4	ccgcgca	2	cgtgacg	20	gccact	1	ggcttgc	2	taggtga	23
aacgaca	7	aggtacg	10	ccgcgcc	31	cgtgact	8	gccacac	1	gggaagc	5	tagtcct	9
aacgcag	5	aggtatg	1	ccgcgcg	4	cgtagat	14	gccacag	2	gggacac	8	tagtcgc	2
aacgccg	76	aggtcac	15	ccgcgct	70	cgtgata	6	gccacca	7	gggaccg	38	tagtcgg	1
aacgcct	7	aggtcag	2	ccgcggg	5	cgtgatc	9	gccaccg	4	gggacct	2	tagtcgt	42
aacgcga	6	aggtccc	1	ccgcgta	25	cgtgcac	61	gccacga	18	gggacgc	37	tagtctg	13
aacgcgc	4	aggtcga	2	ccgcgtc	30	cgtgcca	77	gccacgc	6	gggacgt	8	tagtgcc	9
aacgcta	26	aggtcgc	15	ccgcgtt	1	cgtgccc	4	gccacgg	80	gggacta	2	tagtgcg	1
aacggcc	7	aggtgac	3	ccgctag	12	cgtgcca	5	gccactg	5	gggactc	1	tagtggc	1
aacgggc	8	aggtgat	4	ccgctgc	1	cgtgccg	99	gccacta	1	gggagac	15	tatacgc	11
aacggta	7	aggtggg	2	ccgctgg	66	cgtgccg	2	gccagac	32	gggcaac	7	tatagcg	2
aacgtca	5	aggttac	13	ccgctgt	6	cgtgcgt	71	gccagcc	67	gggcaca	59	tatcacg	38
aactgac	29	aggttgc	3	ccggacg	1	cgtgcca	31	gccagcg	2	gggcacc	10	tatcgca	14
aactggt	11	agttacac	7	ccggact	35	cgtgccc	36	gccagct	1	gggcacg	1	tatcgcg	48
aactgtg	4	agttacgc	8	ccggcag	8	cgtgccg	48	gccagcc	6	gggcagc	7	tatgacg	46
aactgta	5	agttagg	27	ccggcat	13	cgtgggc	8	gccaggg	2	gggcagt	2	tatgccg	29
aactgac	1	agttatgc	2	ccggcca	6	cgtgggg	1	gccaggt	1	gggccac	46	tatgcga	5
aactgcc	2	agttacac	1	ccggccc	19	cgtggtta	37	gccagta	2	gggccca	7	tatgcgc	2
aactgcg	13	agttacc	1	ccggccg	30	cgtggtc	5	gccagtc	46	gggccgc	12	tatgtcg	1
aactggc	5	agttacg	3	ccggcct	55	cgtgtaa	1	gccagtg	1	gggcctc	1	tatgtgc	1
aagacac	1	agttacg	18	ccggcgc	99	cgtttag	2	gccatcg	3	gggcgca	12	tattcgc	9
aagacgc	6	agttacc	1	ccggcgg	5	cgtttca	4	gccatgt	11	gggcgcc	12	tattgcg	1
aagccgc	1	agttccg	90	ccggcgt	45	cgtttcc	4	gccatta	1	gggcgct	1	tcaagcg	40
aagcgac	1	agttgac	1	ccggcta	10	cgtttcg	34	gccacc	1	gggcggc	55	tcacagt	4
aagcgcc	1	agttcgc	62	ccggctg	27	cgtttgc	4	gccact	3	gggcggg	14	tcacca	5
aagcgcg	9	agttcgt	1	ccgggag	2	cgtttgg	1	gccatc	1	gggcgtc	57	tcacctg	2
aagcggc	9	agttcgtt	3	ccgggat	1	cgttttg	10	gccccac	4	gggcgtg	1	tcacctg	1
aagcgtg	1	agttcgtc	55	ccgggct	4	cgttaaa	14	gcccccg	5	gggctac	37	tcacgca	30
aaggagc	1	agttcgtg	2	ccggggt	49	cgttaac	3	gccccct	6	gggctag	6	tcacgcc	57
aaggcaa	4	agttctaa	4	ccgggtc	1	cgttacc	3	gccccgc	5	gggctcc	1	tcacgcg	63
aaggcca	1	agttctac	1	ccgggtg	3	cgttagg	9	gccccgac	8	gggctgc	1	tcacggt	4

Продолжение таблицы 6.

Слово	n	Слово	n	Слово	n	Слово	n	Слово	n	Слово	n	Слово	N
aaggccg	6	agtctag	4	ccggtcc	5	cgtagt	3	gcccgat	3	ggggaca	1	tcacgtg	25
aaggcgc	8	agtctgc	11	ccggtcg	6	cgttcag	79	gcccgcc	10	ggggacg	11	tactgt	6
aaggctg	1	agtctgt	14	ccggtct	9	cgttgac	2	gcccgcg	10	ggggcca	5	tcagagg	1
aagttag	4	agtcttg	1	ccggtgc	41	cgttgcc	5	gcccgct	15	ggggccg	7	tcagatg	1
aagttag	1	agttagc	1	ccggtta	7	cgttggg	2	gcccggc	11	ggggcct	5	tcagccg	30
aagttag	1	agttagc	2	ccgtaca	1	cgttgtc	5	gcccggg	71	ggggcgg	15	tcagcga	26
aagttag	1	agttagg	1	ccgtacc	1	cgttgtg	1	gcccgta	16	ggggggc	4	tcagcgt	77
aagttag	9	agttagc	70	ccgtata	1	cgtttaa	5	gcccgtt	1	gggggtac	20	tcagcta	3
aatcacc	1	agttagc	9	ccgtcca	3	cgtttcg	9	gcccctag	46	gggggtg	7	tcaggac	1
aatcacc	2	agttagc	1	ccgtcta	6	ctaaccg	1	gccccta	1	gggtaac	6	tcaggtg	3
aatcacc	1	agttagc	1	ccgtctg	14	ctaaccg	12	gcccctc	6	gggtacg	6	tcaggtt	1
aatcacc	1	agttagc	109	ccgtgcc	2	ctaaccg	1	gcccctg	1	gggtatg	31	tcagtag	1
aatcacc	17	agttagc	1	ccgtgcg	36	ctaaccg	2	gcccctg	7	gggtcac	5	tcagtag	6
aatcacc	35	agttagc	1	ccgtgct	1	ctaaccg	29	gcccacc	9	gggtccg	8	tcacgc	4
acaacac	17	agttagc	1	ccgtggc	22	ctaaccg	10	gcccacc	56	gggtgcg	26	tcacgc	36
acaacac	1	agttagg	6	ccgtggg	1	ctacacg	2	gcccacc	4	gggtggg	5	tcacgtg	1
acaacac	8	agttagc	13	ccgtgta	21	ctacact	2	gcccacc	43	gggtgtc	5	tccaacc	36
acaacac	2	agttagc	1	ccgttaa	5	ctaccgg	1	gcccata	8	gggtgtg	16	tccaccg	1
acaacac	5	agttagc	3	ccgttag	3	ctacgat	1	gcccata	34	ggtaact	1	tcccacc	1
acaacac	7	agttagc	5	ccgttcg	5	ctacgca	6	gcccata	4	ggtacac	2	tcccacc	5
acaacac	24	agttagc	4	ccgttgc	11	ctacgcg	35	gcccaca	61	ggtaccg	84	tcccccg	1
acaacac	5	agttagc	11	ccgttgg	1	ctacgct	28	gcccacc	1	ggtaccg	2	tccccgc	2
acaacac	4	agttagc	1	ccgttgt	4	ctacgct	1	gcccacc	10	ggtaccg	69	tccccgg	1
acaacac	4	agttagg	1	ccgttta	35	ctactca	5	gcccacc	14	ggtaccg	2	tcccacc	16
acaacac	36	agttagc	1	cctaact	2	ctagaca	5	gcccacc	1	ggtaccg	10	tcccacc	8
acaacac	38	agttagc	1	cctaage	2	ctagacc	1	gcccacc	53	ggtaccg	5	tcccacc	5
acaacac	4	agtttgc	1	cctacac	2	ctagact	1	gcccacc	6	ggtaccg	1	tcctacc	5
acaacac	5	agtttgt	5	cctaccg	5	ctagacc	39	gcccacc	33	ggtaccg	2	tcctacc	7
acaacac	66	ataacac	11	cctaccg	6	ctagacc	4	gcccacc	2	ggtaccg	12	tcctacc	4
acaacac	1	ataacgt	1	cctagac	6	ctagacc	70	gcccacc	79	ggtaccg	5	tccgacc	1
acaacac	2	ataaccg	1	cctagcg	9	ctagacc	37	gcccacc	17	ggtaccg	1	tccgacc	2
acaacac	102	ataaccg	1	cctagtc	18	ctagacc	93	gcccacc	59	ggtaccg	13	tccgacc	4
acaacac	3	ataaccg	2	cctagtg	7	ctagacc	1	gcccacc	16	ggtaccg	6	tccgacc	4
acaacac	6	ataaccg	1	cctatgc	1	ctagacc	8	gcccacc	1	ggtaccg	16	tccgacc	17
acaacac	13	atagccg	1	cctacc	1	ctagacc	7	gcccacc	98	ggtaccg	6	tccgacc	8
acaacac	18	atagcgc	78	cctacc	4	ctagacc	5	gcccacc	83	ggtaccg	24	tccgacc	2
acaacac	1	atagcgt	3	cctacc	4	ctagacc	42	gcccacc	4	ggtaccg	6	tccgacc	7
acaacac	25	atagcgc	11	cctacc	1	ctagacc	3	gcccacc	1	ggtaccg	12	tccgacc	4

Продолжение таблицы 6.

Слово	n	Слово	n	Слово	n	Слово	n	Слово	n	Слово	n	Слово	N
acactac	6	atatcgc	24	cctcagt	1	ctagtca	6	gccgggc	12	ggtcgcg	6	tcctcgc	1
acactag	2	atcaccg	13	cctcatc	2	ctagtcc	1	gccgggg	18	ggtcgtc	43	tcctctc	1
acactca	35	atcacga	2	cctccgt	1	ctagtct	2	gccgggt	60	ggtcggg	15	tcctgfc	1
acactcg	1	atcacgc	4	cctcgac	1	ctatcgc	5	gccggta	3	ggtcggg	1	tcctgtg	1
acactgc	9	atcacgt	16	cctcgca	44	ctatgca	2	gccggtt	1	ggtcgtc	36	tcctcac	6
acactgt	29	atcccgg	1	cctcgct	35	ctatggc	1	gccgtca	3	ggtctag	6	tcctgca	1
acagact	4	atccgca	1	cctcgga	1	ctcaact	1	gccgtcc	81	ggtctca	1	tcctcgc	1
acagaga	5	atccgcg	5	cctcggc	6	ctcacac	2	gccgtct	10	ggtctcg	3	tcctgtc	31
acagatg	1	atcgccc	4	cctcgta	1	ctcacag	2	gccgtgc	2	ggtctgc	6	tcctgtg	5
acagcag	2	atcgctt	1	cctgacc	1	ctcacat	5	gccgtgg	9	ggtgaca	5	tcgacac	7
acagcct	27	atcgcg	60	cctgagt	37	ctcaccg	12	gccgtgt	6	ggtgact	6	tcgaccg	14
acagcga	1	atcgctt	30	cctgcaa	1	ctcacgc	22	gccgtta	4	ggtgagc	2	tcgactg	57
acagcgc	13	atcgctg	6	cctgcag	10	ctcacgg	4	gccgttc	1	ggtgatg	3	tcgagac	20
acagcgg	23	atcgggc	10	cctgccg	62	ctcactc	6	gccgttg	1	ggtgcc	34	tcgagcg	1
acagcgt	74	atcgggc	25	cctgcga	1	ctcagac	5	gcctaac	1	ggtgccg	3	tcgaggg	1
acagctg	1	atcgggg	5	cctgcgc	5	ctcagcg	1	gcctaca	4	ggtgcga	17	tcgacaca	1
acaggcc	2	atcggtg	19	cctgcgg	1	ctcaggc	20	gcctacc	36	ggtgcgc	94	tcgacaga	1
acaggcg	1	atctacg	9	cctgcgt	40	ctcccgc	3	gcctatc	1	ggtggca	7	tcgacagc	76
acagggc	8	atctcgc	1	cctgtctg	8	ctcccgt	1	gcctcag	1	ggtggcc	79	tcgacagg	3
acagtaa	4	atctgca	4	cctggca	3	ctcccta	2	gcctcgc	57	ggtggga	5	tcgacata	1
acagtcc	6	atgacgc	51	cctggcc	6	ctccctg	3	gcctcgg	12	ggtgggc	1	tcgacatg	7
acagtcg	4	atgcaca	16	cctgggt	2	ctccgcg	13	gcctcgt	6	ggtgggg	1	tcgccag	1
acagtga	1	atgact	1	cctgtaa	1	ctccggc	5	gcctctc	5	ggtggta	37	tcgccga	59
acagtgc	5	atgccc	1	cctgtcg	2	ctcgaca	2	gcctgag	28	ggtggtc	1	tcgccgt	1
acagtgg	58	atgccga	1	cctgtct	2	ctcgacc	1	gcctcgc	108	ggtgtcc	1	tcgcctg	1
acagtgt	27	atgccgg	2	cctgtgg	14	ctcgact	4	gcctgct	1	ggtgtct	12	tcgcgac	3
acatcag	5	atgccgt	23	ccttcgc	3	ctcgagg	1	gcctggc	7	ggtgtga	5	tcgcgag	47
acatcgc	7	atcgcaa	2	ccttgag	10	ctcgagt	1	gcctgtg	7	ggtgtgc	2	tcgcgca	23
acatcgg	1	atcgac	43	cgaacgg	7	ctcgcaa	2	gccttaa	1	ggttcac	1	tcgcgcg	51
acatgac	6	atcgca	11	cgaagac	4	ctcgcac	2	gccttat	3	ggttcgc	8	tcgcgct	62
acatgcg	3	atcgccc	4	cgaaggt	1	ctcgcag	41	gccttca	21	ggttgca	9	tcgcggc	1
acatggc	3	atcgcg	18	cgaatgc	16	ctcgcat	4	gcgaacc	2	ggttgcc	1	tcgcggg	1
accacag	63	atcgggc	7	cgacaac	14	ctcgcca	1	gcgaagt	1	ggttgcc	55	tcgcgtg	2
accacgc	40	atcggtt	1	cgacacc	6	ctcgccg	5	gcgaatg	10	gtaacac	42	tcgcgtt	64
accagcc	19	atcgctc	8	cgacacg	24	ctcgcgt	1	gcgacaa	2	gtaacat	2	tcgctaa	4
accagcg	1	atcgctg	10	cgacagt	39	ctcgcta	3	gcgacac	1	gtaacct	9	tcgctct	8
accagta	2	atgctgg	7	cgacata	2	ctcgctc	49	gcgacag	14	gtacacg	67	tcgctga	11
accagtg	29	atggcac	3	cgaccac	6	ctcgctg	10	gcgacat	1	gtacact	52	tcgctgc	49

Продолжение таблицы 6.

Слово	n	Слово	n	Слово	n	Слово	n	Слово	n	Слово	n	Слово	N
accatcg	5	atggcgc	1	cgaccag	4	ctcggac	1	gcgacca	77	gtacaga	1	tcgctgg	18
accatgg	5	atggcgt	1	cgaccga	1	ctcggcc	1	gcgaccg	1	gtacagc	3	tcgctgt	7
accsaca	1	atggctg	3	cgaccgg	13	ctcggcg	17	gcgacct	15	gtacccc	36	tcgcttg	1
accsacg	1	atggctc	2	cgaccgt	31	ctcgggc	1	gcgacga	1	gtaccga	4	tcggact	1
accsagc	2	atgtcat	1	cgaccta	3	ctcgtca	2	gcgacgc	115	gtaccgg	1	tcggagc	2
accsagg	1	atgtcgc	5	cgaccte	46	ctcgtcc	20	gcgacgg	1	gtacgac	39	tcggagg	2
accscac	1	atgtcgt	2	cgacctg	2	ctcgtcg	2	gcgacgt	116	gtacgca	56	tcggcac	2
accscgg	36	atgtcgc	7	cgacgac	80	ctctcac	1	gcgacta	3	gtacgcc	3	tcggccg	55
accsggc	55	atgttgc	7	cgacgca	1	ctctcgc	3	gcgactt	3	gtacgcg	110	tcggcga	6
accsggg	60	atfcgct	1	cgacgcc	16	ctctcgc	22	gcgagca	5	gtacgct	83	tcggcgc	1
accctac	5	attgcga	1	cgacgcg	35	ctctgtc	1	gcgagcc	1	gtacgga	2	tcggcgt	44
accctca	1	caaacgt	54	cgacgct	3	ctgaacg	18	gcgagcg	1	gtacggc	5	tcggctg	2
accctgt	5	caaacgc	1	cgacggc	6	ctgacag	23	gcgagct	8	gtacggg	54	tcgggcg	1
accgacg	27	caacacg	27	cgacgta	2	ctgacct	7	gcgagga	1	gtacgtc	16	tcggggt	1
accgagt	3	caaccgc	51	cgacgtc	46	ctgacgt	1	gcgaggc	43	gtactca	4	tcgggtg	4
accgcat	5	caacgcc	1	cgacgtg	3	ctgactg	5	gcgagta	65	gtactcg	7	tcggtaa	1
accgcca	26	caacgcg	66	cgactaa	12	ctgagcg	5	gcgagtg	52	gtagacg	1	tcggtg	2
accgcca	2	caacgct	38	cgactag	8	ctgaggg	1	gcgataa	1	gtagact	1	tcggtgt	28
accgccc	31	caacggc	2	cgactcg	1	ctgaggt	1	gcgatac	16	gtagcgc	9	tcgtacg	12
accgccc	1	caacggt	7	cgactgc	50	ctgagtc	39	gcgatag	17	gtagcgt	84	tcgtcac	5
accgctg	1	caacgta	8	cgactgg	1	ctgagtg	18	gcgatat	1	gtaggac	2	tcgtcat	1
accggac	5	caacgtg	15	cgagaac	3	ctgatgc	12	gcgatca	41	gtaggca	79	tcgtcct	9
accggcc	2	caacgtt	8	cgagacc	3	ctgatgg	1	gcgatcc	4	gtaggcg	1	tcgtcgc	1
accggcg	5	caagcac	1	cgagacg	3	ctgcaca	75	gcgatcg	7	gtagggc	2	tcgtcgg	1
accggct	26	caagcgc	40	cgagcat	1	ctgcacg	3	gcgatgc	1	gtaggtc	4	tcgtcta	6
accgggc	30	caaggcg	1	cgagcgg	13	ctgcaga	5	gcgatgg	4	gtagtgc	10	tcgtctg	3
accggta	6	caatgtc	1	cgagcgt	15	ctgcagc	11	gcgatgt	20	gtatgac	1	tcgtggc	27
accggtc	13	cacacag	62	cgagctt	1	ctgcagg	3	gcgattt	1	gtatgca	1	tctaacg	1
accgtca	1	cacaccc	7	cgaggac	17	ctgccac	5	gcgcaaa	1	gtatgcg	32	tctagcg	1
accgtcc	8	cacacga	1	cgaggat	3	ctgccag	87	gcgcaca	72	gtatgtc	1	tctcgca	10
accgtgg	5	cacacgc	11	cgaggcc	12	ctgcccc	4	gcgcacc	91	gtattac	8	tctcgcc	3
accgtgt	31	cacacgg	5	cgaggcg	53	ctgcccg	17	gcgcacg	33	gtcaacg	34	tctcggc	1
accgttc	3	cacagat	10	cgaggct	14	ctgcccgc	40	gcgcact	52	gtcaagc	5	tctcgtc	19
accgttt	1	cacagcg	2	cgagggc	2	ctgcccgg	15	gcgcaga	9	gtcaatg	1	tctgaca	1
acctagt	4	cacagga	36	cgagggg	14	ctgccgt	1	gcgcagg	2	gtcacag	8	tctgacg	10
acctcac	4	cacaggc	2	cgaggtg	5	ctgcgaa	7	gcgcata	4	gtcacat	1	tctgcac	11
acctcag	3	cacagtc	26	cgagtcg	1	ctgcgag	16	gcgcatac	99	gtcacct	1	tctgcga	1
acctccc	1	cacagtg	89	cgagtgt	12	ctgcgat	1	gcgcatag	2	gtcacga	22	tctgcgc	19

Продолжение таблицы 6.

Слово	n	Слово	n	Слово	n	Слово	n	Слово	n	Слово	n	Слово	N
acctcgt	2	sacagtt	22	cgagtta	41	ctgcgcg	13	gcgcatt	5	gtcacgc	111	tctgcgg	3
acctgca	5	sacatag	2	cgatacc	1	ctgcgct	2	gcgccac	84	gtcacgg	1	tctgcgt	12
acctgcg	23	sacatga	1	cgatacg	1	ctgcggc	53	gcgccag	40	gtcacgt	1	tctggac	14
acctggt	1	saccasa	1	cgatagc	3	ctggegta	30	gcgccat	12	gtcactc	10	tctggcc	3
acctgtc	18	saccacc	4	cgatac	5	ctgegtc	19	gcgcca	1	gtcagac	44	tctggcg	3
acctteg	24	saccagg	2	cgatccc	1	ctgeggtg	20	gcgccc	87	gtcagca	1	tgaacac	2
acgaact	1	saccata	8	cgatcgc	2	ctggacc	14	gcgccct	1	gtcaggt	64	tgaacgt	4
acgaaa	9	saccatg	2	cgatcgt	6	ctggacg	1	gcgccgc	50	gtcagtc	1	tgaagcg	1
acgacac	84	saccac	12	cgatgcg	3	ctggagg	2	gcgccgg	99	gtcagtg	4	tgaaggc	1
acgacag	4	sacccega	2	cgatggc	9	ctggcac	1	gcgccgt	8	gtcatac	1	tgacacc	5
acgaccg	3	sacccegc	4	cgatgtg	1	ctggcag	58	gcgccta	1	gtcatag	2	tgacacg	1
acgacgc	3	saccceg	61	cgcaaca	9	ctggccc	1	gcgcctc	5	gtcatct	1	tgacagt	1
acgacta	12	saccctg	66	cgcaact	3	ctggccg	13	gcgcctt	20	gtcatgc	11	tgaccac	1
acgactc	1	saccgac	5	cgcaagc	43	ctggcct	7	gcgcgaa	4	gtccaac	16	tgaccgt	17
acgactg	1	saccgat	9	cgcaagt	69	ctggcga	1	gcgcgac	95	gtccacc	1	tgacgac	5
acgaggc	15	saccgcg	19	cgcaatc	8	ctggcgc	44	gcgcgat	44	gtccaga	1	tgacgca	56
acgatcg	5	saccggc	1	cgcacac	10	ctggcgt	4	gcgcgca	53	gtccagc	34	tgacgcg	38
acgatgt	8	saccggg	1	cgcacag	5	ctggtag	5	gcgcgcg	44	gtcccac	16	tgacgct	18
acgcaac	76	saccggt	1	cgcacat	1	ctgggtc	2	gcgcgga	10	gtcccag	1	tgacggc	3
acgcaca	5	saccgta	16	cgcacca	6	ctgggtga	1	gcgcggc	36	gtccccc	11	tgacgta	1
acgcacg	105	sacctgc	5	cgcaccc	46	ctgggtgc	1	gcgcggg	61	gtccccc	3	tgacgtg	1
acgcaga	77	sacctgt	4	cgcaccg	5	ctgtaag	1	gcgcggt	20	gtcccct	1	tgacgtt	5
acgcagc	5	sacgaac	15	cgcacgc	63	ctgtacg	5	gcgcgta	27	gtcccgg	1	tgagacc	1
acgcagg	85	sacgaag	51	cgcacgg	11	ctgtcac	66	gcgcgtc	92	gtcccgt	3	tgagagc	1
acgcagt	46	sacgaca	2	cgcacta	1	ctgtcag	4	gcgcgtg	14	gtcccg	5	tgagcgc	41
acgcata	7	sacgacc	1	cgcactg	23	ctgtccg	16	gcgcgtt	65	gtccgga	2	tgagcgt	1
acgcatic	39	sacgact	13	cgcagac	28	ctgtcgc	7	gcgctaa	27	gtccggc	16	tgagctc	4
acgcatic	22	sacgatg	15	cgcagag	58	ctgtcgg	1	gcgctag	76	gtccgta	67	tgagggc	2
acgcaaa	2	sacgcaa	1	cgcagat	2	ctgtctg	5	gcgctat	4	gtccgtc	2	tgagtgg	10
acgccac	20	sacgcac	72	cgcagcc	59	ctgtgac	7	gcgctca	4	gtccgtt	1	tgatcgc	58
acgccag	3	sacgcag	66	cgcagcg	73	ctgtgat	6	gcgctcc	86	gtcctac	1	tgatgcg	53
acgccc	48	sacgcat	17	cgcaggc	77	ctgtgca	25	gcgctcg	6	gtcctag	1	tgatggc	6
acgccga	30	sacgcc	8	cgcaggg	1	ctgtgcg	1	gcgctct	5	gtcctcg	2	tgcaacc	15
acgccgc	62	sacgccc	4	cgcaggt	19	ctgtggc	6	gcgctgc	31	gtcctga	1	tgcaacg	1
acgccgg	75	sacgccg	51	cgcagtc	78	ctgtggt	11	gcgctgg	65	gtcctgc	1	tgcacac	4
acgccgt	90	sacgcct	1	cgcagtt	2	ctgtgtc	4	gcgctgt	15	gtcctta	4	tgcacag	32
acgcgaa	5	sacgcga	1	cgcatag	13	ctgtgtg	17	gcgctta	50	gtcgaca	2	tgcatat	5
acgcgac	23	sacgcgc	78	cgcatca	3	cttaaac	1	gcgcttc	5	gtcgacc	6	tgaccg	16

Продолжение таблицы 6.

Слово	n	Слово	n	Слово	n	Слово	n	Слово	n	Слово	n	Слово	N
acgcat	47	cacgagg	1	cgcatcc	38	cttaaca	1	gcgcttg	54	gtcgacg	1	tgcaact	5
acgcatg	76	cacgagt	92	cgcatcg	17	cttaagc	9	gcggacc	18	gtcgagt	1	tgcaaga	2
acgcatc	91	cacgcta	69	cgcatga	5	cttaagg	2	gcggagt	7	gtcgcaa	1	tgcaagc	6
acgcatg	102	cacgctg	7	cgcatgc	43	cttacac	8	gcggcac	40	gtcgcat	17	tgcaagt	2
acgcatg	2	cacgctt	30	cgcatgt	20	cttacag	18	gcggcag	10	gtcgcca	16	tgcaactg	8
acgcatg	5	cacggac	4	cgcattc	1	cttaccg	8	gcggcca	20	gtcgccg	6	tgcaagc	52
acgcatg	33	cacggat	1	cgcatcg	6	cttacgc	1	gcggccc	16	gtcgccg	21	tgcaagc	1
acgcatg	3	cacggca	69	cgccaac	48	cttacgt	2	gcggcct	82	gtcgcca	2	tgcaagc	12
acgcatg	58	cacggcg	16	cgccacg	49	cttagcg	13	gcggcgc	86	gtcgcca	102	tgcaagc	3
acgcatg	92	cacggct	1	cgccaga	11	cttagtc	2	gcggcgg	21	gtcgcca	33	tgcaagc	5
acgcatg	77	cacgggg	6	cgccagc	1	cttcagg	20	gcggcgt	88	gtcgcca	75	tgcaagt	3
acgcatg	2	cacggta	3	cgccagg	1	cttcgca	1	gcggcct	3	gtcgcca	24	tgcaagc	21
acgcatg	50	cacggtc	56	cgccagt	38	cttcgcg	3	gcgggca	24	gtcgcca	9	tgcaagc	40
acgcatg	14	cacgtac	1	cgccata	8	cttcgtg	3	gcgggcg	4	gtcgcca	8	tgcaagc	59
acgcatg	10	cacgtag	9	cgccccc	3	cttcgac	39	gcgggct	2	gtcgcca	9	tgcaagc	12
acgcatg	1	cacgtcc	5	cgccccg	11	cttcgag	52	gcgggga	5	gtcgcca	37	tgcaagc	2
acgcatg	5	cacgtcg	2	cgccccg	29	cttcgct	1	gcggggc	20	gtcgcca	1	tgcaagc	3
acgcatg	77	cacgtct	1	cgccccg	5	cttcggg	1	gcggggt	71	gtcgcca	14	tgcaagc	31
acgcatg	5	cacgtga	13	cgcccta	1	cttcggc	1	gcgggtg	28	gtcgcca	16	tgcaagc	1
acgcatg	6	cacgtta	6	cgccctc	4	cttcggg	5	gcggtcg	12	gtcgcca	18	tgcaagc	9
acgcatg	4	cacgttc	1	cgccctg	2	cttcgct	1	gcggctt	22	gtcgcca	9	tgcaagc	1
acgcatg	56	cacgttg	3	cgccgac	6	cttcgtg	1	gcgggtc	14	gtcgcca	5	tgcaagc	1
acgcatg	1	caactac	2	cgccgag	56	gaacagc	1	gcgggta	1	gtcgcca	1	tgcaagc	96
acgcatg	4	caactaca	2	cgccgat	16	gaacacg	19	gcgggtg	8	gtcgcca	67	tgcaagc	3
acgcatg	1	caactacc	1	cgccgca	18	gaacgat	5	gcgtaca	31	gtcgcca	42	tgcaagc	13
acgcatg	41	caactacg	6	cgccgcc	84	gaacgcc	1	gcgtacc	10	gtcgcca	17	tgcaagc	1
acgcatg	6	caactagg	2	cgccgcg	36	gaacggt	3	gcgtacg	62	gtcgcca	1	tgcaagc	53
acgcatg	6	caactagt	3	cgccgga	47	gaagccg	1	gcgtact	1	gtcgcca	1	tgcaagc	19
acgcatg	16	caactcag	2	cgccggc	5	gaagcgt	11	gcgtage	1	gtcgcca	1	tgcaagc	15
acgcatg	18	caactccc	8	cgccggg	28	gaagggc	1	gcgtata	7	gtcgcca	1	tgcaagc	17
acgcatg	3	caactccg	2	cgccgta	1	gaagtcc	1	gcgtatc	11	gtcgcca	50	tgcaagc	23
acgcatg	7	caactcgc	46	cgccgtc	14	gaatgcg	3	gcgtcaa	24	gtcgcca	4	tgcaagc	1
acgcatg	4	caactcgg	4	cgccgtg	3	gacaacg	18	gcgtcac	89	gtcgcca	5	tgcaagc	1
acgcatg	18	caactctc	19	cgccgtt	4	gacaagc	3	gcgtcag	5	gtcgcca	1	tgcaagc	1
acgcatg	9	caactgcc	5	cgccctaa	2	gacacac	2	gcgtcat	10	gtcgcca	16	tgcaagc	52
acgcatg	36	caactgcg	1	cgccctat	2	gacaacc	10	gcgtcca	1	gtcgcca	1	tgcaagc	1
acgcatg	9	caactggc	8	cgccctcc	1	gacaccg	5	gcgtccc	8	gtcgcca	5	tgcaagc	96
acgcatg	1	caactggt	6	cgccctcg	27	gacacgc	1	gcgtccg	42	gtcgcca	1	tgcaagc	53

Продолжение таблицы 6.

Слово	n	Слово	n	Слово	n	Слово	n	Слово	n	Слово	n	Слово	N
acggtac	1	caactgtg	32	cgcctgt	5	gacacgt	2	gcgtcct	1	gtctgac	5	tcgagg	57
acggtca	14	caacttac	1	cgcctta	3	gacacta	3	gcgtcga	2	gtctgca	17	tcgagt	1
acggtcg	2	caactcg	4	cgccttg	5	gacactt	8	gcgtcgc	95	gtctgcc	7	tcgatc	8
acggtct	3	cagacac	16	cgcgaaa	1	gacagac	18	gcgtcgg	45	gtctgcg	9	tcgcaa	2
acggtga	1	cagacag	9	cgcgaag	47	gacagcg	9	gcgtcgt	2	gtctggc	10	tcgcac	37
acggtgg	1	cagaccc	4	cgcgaca	25	gacaggt	1	gcgtcta	4	gtctggg	1	tcgcat	23
acggttg	1	cagacga	3	cgcgacc	1	gacagta	3	gcgtctc	1	gtcttag	1	tcgccc	46
acgtaac	6	cagacgc	5	cgcgacg	101	gacagtc	1	gcgtctt	8	gtcttcg	1	tcgcga	48
acgtaag	1	cagacgg	11	cgcgact	79	gacagtg	6	gcgtgaa	1	gtcttta	8	tcgcgc	49
acgtacg	32	cagacgt	1	cgcgagt	69	gaccaca	5	gcgtgac	82	gtgaagc	3	tcgcgg	9
acgtagc	7	cagacta	2	cgcgata	4	gaccacg	3	gcgtgag	1	gtgaagg	4	tcgcgt	8
acgtagt	11	cagactt	54	cgcgatg	2	gaccact	12	gcgtgat	15	gtgaaa	3	tcgcta	1
acgtcac	59	cagagag	17	cgcgatt	2	gaccaga	36	gcgtgcc	3	gtgacac	7	tcgctg	87
acgtcag	2	cagagcg	1	cgcgcaa	26	gaccagt	5	gcgtgcg	81	gtgacag	13	tcggca	15
acgtccg	6	cagaggc	2	cgcgcac	97	gaccceg	16	gcgtgct	21	gtgacgc	24	tcggcc	85
acgtcct	1	cagagtg	6	cgcgcag	36	gaccctg	1	gcgtggc	39	gtgacgt	106	tcggcg	39
acgtcgc	56	cagatta	1	cgcgcat	8	gaccgca	35	gcgtggg	2	gtgacta	8	tcgggg	3
acgtcgg	59	cagaca	64	cgcgcca	28	gaccgcc	1	gcgtggt	59	gtgagag	1	tcggtc	21
acgtcta	1	cagcacg	1	cgcgccg	65	gaccggc	1	gcgtgta	2	gtgagca	1	tcgtac	28
acgtctc	1	cagcact	1	cgcgcct	59	gaccggt	13	gcgtgtc	1	gtgagcg	86	tcgtca	39
acgtgac	2	cagccct	2	cgcgcga	7	gaccgta	3	gcgtgtg	5	gtgaggt	13	tcgtcg	1
acgtgat	2	cagccgg	26	cgcgcgc	60	gaccgtc	51	gcgtgtt	59	gtgagtg	6	tcgtga	8
acgtgcc	8	cagccgt	17	cgcgcgg	23	gaccgtg	57	gcgttac	5	gtgatgc	1	tcgtgc	14
acgtgcg	22	cagcctc	1	cgcgcgt	71	gacctag	2	gcgttgc	26	gtgatta	3	tcgtgg	60
acgtggg	13	cagcctt	1	cgcgcta	2	gacctgc	6	gcgttgg	8	gtgcaac	3	tcgtta	5
acgtgtc	9	cagcgaa	3	cgcgctc	3	gacctgg	2	gcgttgt	1	gtgcaag	12	tgctacg	1
acgtgtg	1	cagcgac	10	cgcgctg	78	gaccttg	1	gcgtttc	1	gtgcaca	10	tgctagc	1
acgtgtt	2	cagcgag	74	cgcgctt	65	gacgaac	6	gcgtttg	1	gtgcacc	37	tgctcac	54
acgttaa	5	cagcgat	4	cgcggaa	1	gacgaca	6	gcgtttt	3	gtgcacg	79	tgctcag	1
acgttac	1	cagcgca	79	cgcggag	10	gacgacc	1	gctaaca	1	gtgcact	39	tgctcgc	1
acgttag	37	cagcgcg	53	cgcggat	1	gacgacg	58	gctaagt	1	gtgccac	18	tgctctg	1
acgttat	11	cagcgct	54	cgcggca	5	gacgact	29	gctaccc	3	gtgcccg	20	tgctgtc	1
acgttca	4	cagcggc	2	cgcggcc	5	gacgcaa	1	gctacct	1	gtgccct	42	tgctgtg	6
acgttgc	2	cagcggg	1	cgcggcg	106	gacgcag	72	gctacgc	24	gtgccga	2	tgacac	26
actaacg	4	cagcgta	10	cgcggga	14	gacgcat	22	gctacgt	5	gtgccgc	91	tgacgc	1
actacac	1	cagcgtc	68	cgcgggc	7	gacgcca	2	gctagca	1	gtgcctc	16	tgacgg	6
actaccg	1	cagcgtg	17	cgcgggg	64	gacgccc	53	gctagcc	21	gtgcgaa	1	tgactg	6
actacgt	6	cagctgg	2	cgcgggt	14	gacgccg	37	gctagcg	97	gtgcgac	55	tgagcc	8

Продолжение таблицы 6.

Слово	n	Слово	n	Слово	n	Слово	n	Слово	n	Слово	n	Слово	N
actaggc	11	cagctgt	5	cgcggtc	3	gacgcga	3	gctagtc	7	gtgcgag	2	tggaggc	2
actaggt	1	caggaca	51	cgcggtg	1	gacgcgc	77	gctcaca	5	gtgcgca	67	tggaggg	1
actagtg	5	caggacg	3	cgcggtt	25	gacgcgg	23	gctcacc	3	gtgcgcc	107	tggatgt	8
actatgc	1	caggact	9	cgcgtag	32	gacgcgt	31	gctcata	2	gtgcgcg	61	tggcacg	16
actcacg	1	caggaga	5	cgcgtag	14	gacgctc	48	gctcatc	1	gtgcgct	81	tggccac	1
actcaga	2	caggcaa	1	cgcgtag	2	gacgctg	4	gctccga	4	gtgcgtc	4	tggccag	4
actcagg	6	caggcag	2	cgcgtag	29	gacggac	78	gctccgg	1	gtgcgtg	91	tggcccc	8
actcagt	2	caggcct	3	cgcgtag	60	gacggca	64	gctcctg	5	gtgctag	1	tggccgc	86
actccgc	2	cagggca	1	cgcgtag	80	gacggcc	1	gctcgac	2	gtgctcc	1	tggccgg	2
actccgg	8	cagggcg	2	cgcgtag	1	gacggcg	3	gctcgca	14	gtgctgg	26	tggccgt	1
actcgac	1	cagggggt	1	cgcgtag	12	gacggct	76	gctcgcg	1	gtgctta	26	tggcctc	1
actcgca	8	cagggtc	2	cgcgtag	5	gacggtc	86	gctcgct	7	gtggact	1	tggcctg	1
actcgcg	55	cagggtg	56	cgcgtag	1	gacggtt	3	gctcgga	1	gtggcac	9	tggcctt	4
actcggt	4	caggtag	1	cgcgtag	29	gacgtac	17	gctcggc	1	gtggcag	3	tggcgac	5
actcgtc	31	caggtag	15	cgcgtag	53	gacgtag	1	gctctag	1	gtggcca	22	tggcgag	25
actctcc	5	cagtaac	11	cgtaaac	9	gacgtcc	46	gctctct	1	gtggccc	13	tggcgcc	13
actctcg	1	cagtact	1	cgtaaac	5	gacgtcg	26	gctctgc	6	gtggccg	5	tggcgcg	27
actgacc	5	cagtagg	1	cgtaaac	1	gacgtct	19	gctgaac	38	gtggcct	12	tggcgct	51
actgacg	18	cagtatg	2	cgtagaa	2	gacgtgc	6	gctgact	47	gtggcgc	114	tggcggg	23
actgagt	31	cagtcca	4	cgtagac	92	gacgtgt	16	gctgaga	1	gtggcgt	1	tggcggt	9
actgcca	2	cagtccg	15	cgtagag	12	gacgtta	28	gctgcac	26	gtgggaa	3	tggcgtc	8
actgccc	41	cagtccct	1	cgtagat	1	gactaac	3	gctgcca	7	gtgggac	1	tggcggt	15
actgcgt	1	cagtcca	5	cgtagat	9	gactaca	3	gctgccc	79	gtgggca	4	tggcggt	21
actgctc	1	cagtccg	5	cgtagat	71	gactagc	8	gctgccc	1	gtgggcg	1	tggcggt	10
actggca	70	cagtcta	19	cgtagca	1	gactagg	4	gctgcga	50	gtgggct	1	tgggacg	1
actggcc	3	cagtctc	4	cgtagcg	2	gactagt	12	gctgcgg	1	gtggggc	5	tgggacg	1
actggcg	28	cagtctg	2	cgtagcg	5	gactcag	8	gctggac	82	gtggggg	4	tggggac	7
actggct	5	cagtcaa	1	cgtagcg	5	gactccg	3	gctggca	1	gtggggg	1	tggggcc	5
actgggt	11	cagtccg	84	cgtagca	54	gactcgc	5	gctggct	62	gtgggtg	3	tggggta	4
actggtc	7	cagtctc	1	cgtagcc	6	gactctg	2	gctggga	27	gtggtag	9	tgggtac	4
actggtg	2	cagtgtt	57	cgtagcg	92	gactgac	1	gctgggg	5	gtggtag	5	tgggtgc	5
actgtca	4	cagttag	1	cgtagga	14	gactgcc	5	gctggta	3	gtggtag	56	tgggtgt	1
actgtcg	66	catagcg	1	cgtaggt	1	gactgcg	59	gctgtaa	9	gtgtagg	11	tggtagg	7
actgtgc	1	catagcg	3	cgtagta	29	gactgct	6	gctgtac	1	gtgtacg	10	tggtagc	82
actgtgg	27	catagcc	6	cgtagtc	14	gactggc	21	gctgtat	3	gtgtagc	1	tggtagc	13
actgtgt	19	catagcg	3	cgtagtg	18	gactggg	1	gctgtcc	16	gtgtagg	2	tggtagc	41
actgtta	1	catagca	2	cgtagtt	4	gactgtg	25	gctgtga	1	gtgtagt	1	tggtagc	55
acttaag	4	catagct	1	cgtagta	4	gagaact	1	gcttaaa	24	gtgtcca	26	tggtagc	53

Продолжение таблицы 6.

Слово	n	Слово	n	Слово	n	Слово	n	Слово	n	Слово	n	Слово	N
acttccc	1	catcgcc	5	cgcttga	7	gagaccc	1	gcttaag	9	gtgtccc	2	tggttgc	19
acttgcc	1	catcgcg	11	cgcttgc	4	gagaccg	2	gcttaca	1	gtgtcct	1	tgtaagc	1
acttgcg	86	catcgct	8	cggaaqt	7	gagacct	1	gcttacg	1	gtgtcga	6	tgtaagg	1
agaacgc	1	catcgtc	1	cggacag	1	gagacgc	73	gcttgcg	54	gtgtcgc	5	tgtaccg	3
agacacc	11	catcgtg	2	cggacat	4	gagacgg	1	gcttggc	1	gtgtcgt	12	tgtacgc	78
agaccac	3	catctgc	2	cggaccc	6	gagactg	4	gcttgtc	8	gtgtctg	13	tgtactc	1
agaccag	14	catgaca	10	cggaccg	2	gagagtg	50	ggaagcc	15	gtgtctt	15	tgtagcc	1
agacccc	1	catgcga	6	cggacga	1	gagccag	1	ggacaag	2	gtgtgag	6	tgtaggc	2
agaccgc	83	catgcgc	3	cggacgc	65	gagcccg	1	ggacacg	17	gtgtgca	7	tgtagtc	1
agaccgg	2	catgcgg	5	cggacgt	4	gagccgc	1	ggacact	62	gtgtgcc	29	gtgcaca	1
agacctg	1	catgcgt	31	cggactc	1	gagcgag	7	ggacagc	14	gtgtgcg	21	gtgcacg	5
agacgac	2	catgtgc	2	cggactg	64	gagcgat	6	ggacccc	14	gtgtggc	1	gtgcagg	4
agacgca	40	ccaacag	9	cggagcg	10	gagcgcc	10	ggacccg	1	gtgtgtc	3	gtgcccc	9
agacgcc	22	ccaacgc	58	cggaggt	7	gagcggc	48	ggaccgc	7	gtgttac	6	gtgccgg	1
agacgcg	2	ccaacgg	1	cggagtc	5	gagcggg	1	ggacgac	35	gtgttag	12	gtgccgt	2
agacggc	9	ccaacgt	3	cggagtg	2	gagcgta	1	ggacgag	6	gtgttcc	1	gtgccta	1
agacggg	1	ccacacg	4	cggatgt	1	gagcgtc	88	ggacgcc	4	gtgttga	1	gtgcctg	1
agacgtc	17	ccacagt	66	cggcaac	9	gaggaca	4	ggacgcg	80	gtgttgc	6	gtgcgac	9
agacgtg	5	ccaccag	10	cggcaag	15	gaggacg	2	ggacgct	7	gttaacg	33	gtgcgag	1
agactag	2	ccaccat	1	cggcaca	60	gaggcac	2	ggacggc	90	gttaact	1	gtgcgca	6
agactat	1	ccaccga	3	cggcacc	2	gaggccc	2	ggacggt	5	gttacac	4	gtgcgcc	5
agactcg	5	ccaccgt	11	cggcact	1	gaggccg	5	ggacgta	1	gttaccg	61	gtgcgcg	47
agactgt	4	ccacgac	3	cggcagc	9	gaggcga	6	ggacgtc	25	gttacgc	1	gtgcgct	6
agagcga	15	ccacgca	62	cggcagg	55	gaggcgg	2	ggacgtg	3	gttagac	6	gtgcggt	2
agagctg	2	ccacgcc	3	cggcata	13	gaggcgt	1	ggactag	9	gttagcg	25	gtgcgta	3
agaggtc	8	ccacggc	5	cggcatc	2	gaggctg	3	ggactgc	10	gttaggc	56	gtgcgtg	36
agagtgg	1	ccacgft	5	cggccac	13	gagggag	1	ggactgt	14	gttagta	1	gtgtctc	4
agatgac	3	ccactac	1	cggccag	2	gagggca	2	ggagaac	1	gttagtc	2	gttgacc	1
agatggg	1	ccactag	6	cggcccc	31	gagggcc	1	ggagcac	4	gttcacc	1	gttgacg	5
agcaacg	1	ccactgt	2	cggccccg	14	gagggcg	47	ggagccc	1	gttcagc	1	gttgagc	1
agcaagc	10	ccagacc	7	cggccct	5	gaggggt	29	ggagcct	5	gttcgcg	9	gttgcac	26
agcacac	1	ccagacg	7	cggccga	35	gagggtc	1	ggagcga	1	gttcgct	5	gttgcac	16
agcaccg	5	ccagcac	88	cggccgg	29	gaggtca	1	ggagcgc	84	gttcggc	5	gttgccc	2
agcacgc	17	ccagcca	1	cggccta	60	gaggtcg	70	ggaggac	1	gttcgtg	42	gttgccg	3
agcacgg	1	ccagccc	1	cggcctc	15	gaggtga	4	ggaggct	1	gttcctg	1	gttgcca	9
agcactt	1	ccagccg	5	cggcgca	16	gaggtgt	4	ggagggc	1	gttgatc	1	gttgccg	15
agcagac	2	ccagcct	6	cggcgcc	72	gagtcgt	2	ggagtca	2	gttgcac	5	gttgctg	49
agcagcg	10	ccagcga	15	cggcgcg	86	gagtcta	15	ggagtgc	14	gttgccg	1	gttgccc	16

Продолжение таблицы 6.

Слово	n	Слово	n	Слово	n	Слово	n	Слово	n	Слово	n	Слово	N
agcaggc	1	ccagcgc	74	cggcgct	1	gagtgac	23	ggagtgg	2	gttgccg	2	tgtggcg	4
agccagc	69	ccagcgg	29	cggcgga	4	gagtgag	4	ggatgcg	4	gttgctt	9	tgtggtc	2
agccagt	44	ccagctg	2	cggcgta	3	gagtgcc	38	ggatgtg	5	gttgctt	35	tgtgtag	1
agcccag	1	ccagctt	6	cggcgtc	62	gagtgta	7	ggcaaag	1	gttggcc	3	tgtgtcc	6
agcccga	6	ccaggac	1	cggcggt	85	gagtgtt	9	ggcaaca	1	gttgccg	6	tgtgtgc	7
agcccgc	18	ccaggct	3	cggcggt	69	gatcacg	1	ggcaacc	4	gttgctt	4	tgttgtt	3
agcccgg	9	ccagggt	1	cggctac	1	gatcccc	1	ggcaacg	109	gtttaag	10	tgtgttc	1
agccctc	1	ccaggtg	1	cggctag	1	gatcgca	1	ggcaagg	23	gtttgcg	1	tgttacc	1
agccgat	5	ccagtct	1	cggctga	16	gatcggc	2	ggcaatg	1	taacacc	2	tgttagc	9
agccgcc	1	ccagtgt	1	cggctgt	1	gatgagg	1	ggcacac	33	taacact	2	tgttcac	15
agccggg	2	ccatcac	1	cggctta	1	gatgccg	17	ggcacca	7	taaccag	10	tgttcgg	1
agccggt	48	cccaccg	3	cggcttc	1	gatcgca	14	ggcaccg	5	taacctg	6	tgttctc	1
agccgtc	27	cccacgc	24	cgggacg	16	gatcgcc	78	ggcacgc	5	taacggc	7	tgttgcc	17
agccgtg	1	cccacgt	6	cgggcaa	16	gatcgtt	1	ggcacgg	4	taacggg	35	tgttgcg	16
agcctac	9	cccactc	1	cgggcac	69	gatggcg	2	ggcacgt	16	taacgtc	2	tgttggc	1
agcctag	8	cccagcg	1	cgggcag	50	gatgtgc	45	ggcactc	1	taacgtg	13	ttaacct	1
agcctcg	22	cccagtc	4	cgggccg	7	gatgtgt	1	ggcactt	11	taacgtt	1	ttaacgg	6
agcgaca	1	cccagtg	1	cgggcgc	33	gattgcg	2	ggcagac	77	taagcac	1	ttaacgt	2
agcgact	10	cccacag	3	cgggcgt	69	gcaaacg	55	ggcagcc	73	taagcgc	2	ttaaccc	20
agcgagc	7	ccccctg	1	cgggcta	5	gcaacac	5	ggcagcg	10	taagcgt	6	ttaacgg	15
agcgagg	16	ccccgac	5	cgggctc	15	gcaacca	4	ggcagct	1	taaggcg	1	ttaatgt	3
agcgatg	56	ccccgca	5	cggggac	38	gcaaccg	19	ggcagga	4	taaggtc	2	ttacacg	9
agcgcac	82	ccccgcc	1	cggggat	1	gcaacct	12	ggcaggg	4	taagtgc	1	ttacagc	1
agcgcat	5	ccccgcg	4	cggggca	4	gcaacgc	20	ggcaggt	2	taagtct	1	ttacagt	5
agcgcca	46	ccccgct	5	cggggcc	10	gcaacgg	11	ggcatac	1	taagtgc	1	ttaccgg	12
agcgccc	2	ccccggc	23	cggggcg	24	gcaacgt	58	ggcatcg	5	taatcgg	1	ttacgca	1
agcgccg	2	ccccgtg	1	cgggggc	1	gcaactg	1	ggccaac	2	taatcgg	1	ttacgct	1
agcgcga	25	cccctcg	2	cggggtc	5	gcaagcc	11	ggccaca	8	tacacag	10	ttagacg	58
agcgcgg	76	cccctgc	1	cggggtg	5	gcacaaa	7	ggccacc	68	tacacgc	63	ttagact	2
agcgcgt	17	cccgaca	5	cgggtca	3	gcacaac	1	ggccacg	8	tacacgg	24	ttagcga	10
agcgcta	12	cccagcg	50	cgggtcg	1	gcacaca	1	ggccact	48	tacactc	2	ttagcgc	19
agcgctc	8	cccgact	5	cgggtgg	1	gcacacc	61	ggccaga	17	tacactg	1	ttagcgt	72
agcgctg	90	cccgatg	16	cgggtgt	9	gcacacg	2	ggccagg	3	tacagcg	31	ttaggcg	8
agcggac	1	cccgcca	8	cggtaac	1	gcacact	1	ggccccac	1	tacaggg	5	ttagggc	1
agcggag	1	cccggcc	1	cggtaac	1	gcacaga	1	ggccccca	6	taccacg	1	ttagtgc	2
agcggcc	20	cccggcg	10	cggtcac	42	gcacagc	1	ggccccg	26	taccagc	1	ttatcgc	39
agcgggc	1	cccgcga	4	cggtcat	1	gcacagg	2	ggccccg	1	taccgac	9	ttatgtg	1
agcgggg	1	cccgcgc	3	cggtccc	69	gcacatc	45	ggcccgt	14	taccgcc	4	ttacagt	5

Продолжение таблицы 6.

Слово	n	Слово	n	Слово	n	Слово	n	Слово	n	Слово	n	Слово	N
agcggtc	5	cccgcgt	26	cggtccg	1	gcacatg	9	ggccta	1	taccggc	14	ttactg	1
agcgtac	47	cccgtg	3	cggtcga	3	gcaccac	1	ggcctc	1	taccggg	11	ttcagcg	6
agcgtag	1	cccggac	1	cggtcgc	5	gcaccat	5	ggccgag	1	tacctca	1	ttcaggc	1
agcgtat	1	cccggca	7	cggtcta	1	gcacccc	1	ggccgat	2	tacctgc	1	ttccgcg	1
agcgtca	15	cccggcc	2	cggtctc	30	gcacccg	1	ggccgca	78	tacctgg	1	ttcctgc	1
agcgtcc	6	cccggcg	7	cggtgac	1	gcaccct	1	ggccgcc	4	tacgagg	2	ttcgacg	6
agcgtcg	3	cccggct	7	cggtgag	19	gcaccga	1	ggccgcg	6	tacgcaa	5	ttcgcac	1
agcgtgc	83	cccgggt	28	cggtgca	29	gcaccgc	19	ggccgct	24	tacgcag	46	ttcgcag	9
agcgtgt	48	cccggtc	5	cggtgct	1	gcaccgg	31	ggccggc	92	tacgcat	1	ttcgcgc	26
agcgttg	3	cccgtat	4	cggtggc	4	gcaccgt	5	ggccggt	8	tacgcca	2	ttcgctg	7
agctacc	1	cccgtct	1	cggtggg	9	gcacgac	7	ggccggt	4	tacgccg	34	ttcggcc	4
agctcaa	1	cccgtta	16	cggtgta	2	gcacgat	1	ggcctca	1	tacgcga	2	ttctgcg	3
agctcgc	15	ccctaag	2	cggtgtt	45	gcacgca	27	ggcctcg	61	tacgcgc	54	ttgaccg	4
agctcgg	1	ccctacg	5	cggttac	3	gcacgcc	8	ggcctct	1	tacgcgg	38	ttgacgc	1
agctgcc	3	ccctact	1	cggttag	29	gcacgcg	10	ggcctgc	1	tacgcgt	6	ttgccag	1
agctgta	9	ccctagg	5	cggttgc	5	gcacgct	59	ggccttc	5	tacgcta	22	ttgccgc	3
agctgtc	16	ccctcaa	1	cggttgg	4	gcacgga	47	ggcgacc	7	tacgctc	3	ttgccgg	2
aggacac	1	ccctcag	6	cggttgt	1	gcacggg	1	ggcgacg	6	tacgctg	29	ttgccgt	8
aggacga	26	ccctgac	5	cgtaac	5	gcacggt	1	ggcgaga	1	tacggag	1	ttgcgac	13
aggact	3	ccctgcg	5	cgtacac	7	gcacgtc	3	ggcgagt	4	tacggca	4	ttgcgca	3
aggactc	1	ccctggc	3	cgtaccg	1	gcacggt	1	ggcgcac	42	tacggcc	4	ttgcgcg	10
aggactg	1	ccctggt	3	cgtacct	2	gcactaa	1	ggcgcat	6	tacggtc	1	ttggggc	65
aggactt	2	ccctgtc	1	cgtacga	8	gcactac	8	ggcgcca	13	tacggtg	1	ttggccg	27
aggagac	4	ccctgtg	4	cgtacgc	106	gcactcc	11	ggcgccc	11	tacgtag	1	ttggcgg	4
aggagcg	1	ccctteg	1	cgtacgg	12	gcactcg	1	ggcgccg	80	tacgtta	1	ttgggtg	2
aggatac	1	ccgaagg	5	cgtacgt	12	gcactga	5	ggcgcct	65	tacgttg	2	ttggtgc	1
aggcaac	60	ccgacaa	4	cgtagcc	5	gcactgc	6	ggcgcgc	4	tactcag	2	ttgtcgc	8
aggcaag	1	ccgacag	1	cgtagtg	2	gcactgg	1	ggcgcgg	34	tactcgc	52	ttgtcgt	3
aggcaca	1	ccgacat	5	cgtatcg	9	gcagaac	5	ggcgcgt	62	tactgac	1	ttgtctg	4
aggcacg	1	ccgaccg	1	cgtcaac	2	gcagaca	16	ggcgctc	7	tactgcc	1	tttaagc	4
aggcatg	2	ccgacga	52	cgtcacg	100	gcagacc	31	ggcgctt	1	tactggc	3	tttagcg	15
aggccac	12	ccgacgc	23	cgtcact	13	gcagacg	76	ggcggag	18	tagacgg	5	tttgccg	4
aggccag	26	ccgacgt	46	cgtcaga	3	gcagact	43	ggcggcg	61	tagactg	1	tttgccc	1
aggccgc	86	ccgacta	45	cgtcagt	8	gcagagc	8	ggcgggc	11	tagacc	1	tttgccg	1
aggccgg	75	ccgagca	1	cgtcata	4	gcagatg	1	ggcgggg	1	tagcccg	8		
aggccgt	63	ccgagcg	27	cgtcatg	63	gcagcac	2	ggcgggt	66	tagccgc	8		
aggcctg	2	ccgagct	1	cgtcccc	1	gcagccc	13	ggcggtt	4	tagccgg	1		
aggcgac	30	ccgaggc	13	cgtcccg	10	gcagcga	14	ggcgtag	1	tagccta	47		

Окончание таблицы 6.

Слово	n	Слово	n	Слово	n	Слово	n	Слово	n	Слово	n	Слово	N
aggcgcc	34	ccgaggt	4	cgtccgg	8	gcagcgc	48	ggcgtca	4	tagcctg	1		

Результаты, представленные в таблице 7, показывают крайне медленный рост длины потерянных слов при увеличении длины последовательности.

Таблица 7–перепись числа наименьших потерянных слов

ID	Организм	Длина генома	Длина наименьшего слова	Количество наименьших слов
FR775227	Salmonella enterica subsp. Mitochondrion	17569	5	2
HQ184045	Bos taurus isolate Mcg375 mitochondrion	16340	5	13
JF727176	Pan troglodytes isolate Flo mitochondrion	16557	5	13
KC469587	Sus scrofa domesticus breed pietrain mitochondrion	16612	5	12
KM061558	Canis lupus familiaris isolate Cf stp64 mitochondrion	16730	5	13
AY217738	Eimeria tenella	34750	5	4
AY945289	Fusarium oxysporum strain F11 mitochondrion	34477	5	2
DQ508940	Debaryomyces hansenii mitochondrion	29462	5	2
DQ642846	Plasmodium falciparum HB3	29529	5	92
JQ864234	Candida albicans strain L296 mitochondrion	33631	5	8
EU651892	Hemiselmis andersenii strain CCMP 644mitochondrion	60553	5	2
FR775213	Salmonella enterica subsp. enterica serovar Weltevreden	64694	6	33
FR775245	Salmonella enterica subsp. enterica serovar Weltevreden	63517	5	4
HG004427	Campylobacter fetus subsp. Venerealis	61142	5	9
KF285530	Ostreococcus tauri isolate RCC1123 chloroplast	67681	6	13
AB042240	Triticum aestivum chloroplast	134545	7	399
CP00224	Candidatus Tremblaya princeps PCIT	138927	6	3
FR775217	Salmonella enterica subsp. enterica serovar Weltevreden	131230	6	19
JN861109	Oryza sativa Indica Group cultivar Hassawi chloroplast	134448	7	361
X86563	Zea mays complete chloroplast genome	140384	7	351
CP000351	Leptospira borgpetersenii, chromosome 2	299762	7	85
CP002163	Candidatus Sulcia muelleri CARI,	276511	6	99
CP007234	Candidatus Sulcia muelleri strain TETUND	270029	6	81
FR775191	Salmonella enterica	227697	6	16

FR775236	Salmonella enterica serovar Weltevreden	253936	6	9
AY506529	Zea mays strain NB mitochondrion	569630	8	1059
CP002243	Candidatus Moranelia endobia PCIT	538294	7	6

Окончание таблицы 7.

ID	Организм	Длина генома	Длина наименьшего слова	Количество наименьших слов
CP003000	Blattabacterium	587248	6	10
CP003771	Mycoplasma genitalium M6282	579504	6	11
CP006771	Mycoplasma parvum str. Indiana	564395	6	8
AL954800	H. sapiens, chromosome 14	87191216	9	110
CM000265	H. sapiens, chromosome 14	87316725	9	110
CM000856	Callithrix jacchus, chromosome 1	210400635	9	3
CM000878	Callithrix jacchus, chromosome X	142054208	9	62
CM000879	Callithrix jacchus, chromosome Y	2853901	7	4
CM000001	Canis lupus familiaris, chromosome 1	122678785	9	4
CM000002	Canis lupus familiaris, chromosome 2	85426708	9	11
CM000003	Canis lupus familiaris, chromosome 3	91889043	9	41
CM000004	Canis lupus familiaris, chromosome 4	88276631	9	63
CM000356	Anopheles gambiae str. PEST chromosome 2L	49364325	10	16
AE005172	Arabidopsis thaliana chromosome 1	14221815	8	1
AP011156	Acetobacter pasteurianus IFO 3283-32	2904642	6	490

Для бактерий показано сильное подобие потерянных слов их сайтам рестрикции. Чтобы доказать это на примере бактерии Acetobacter pasteurianus, приведу сайт рестрикции: 5'...G/TGCAC...3', [7] и участок списка потерянных слов, начинающихся с тройки гуанин-тимин-гуанин: gtagcc gtagcg gtagct gtagcc gtagga gtaggt gtagc gtagct gtagta, и с тройки цитозин-аденин-цитозин: cacacc cacccc cacccg cacctc cacgac cacgga cactag cactcc cactga cactta cacttg. Как видно, с учетом замен, вставок и удалений, они сильно похожи.

Распределение GC-контента некоторых хромосом в виде гистограмм представлены на рисунках 1-4. По оси X гистограмм представлен собственно GC-контент, а по оси Y количество потерянных слов с указанным GC-контентом.

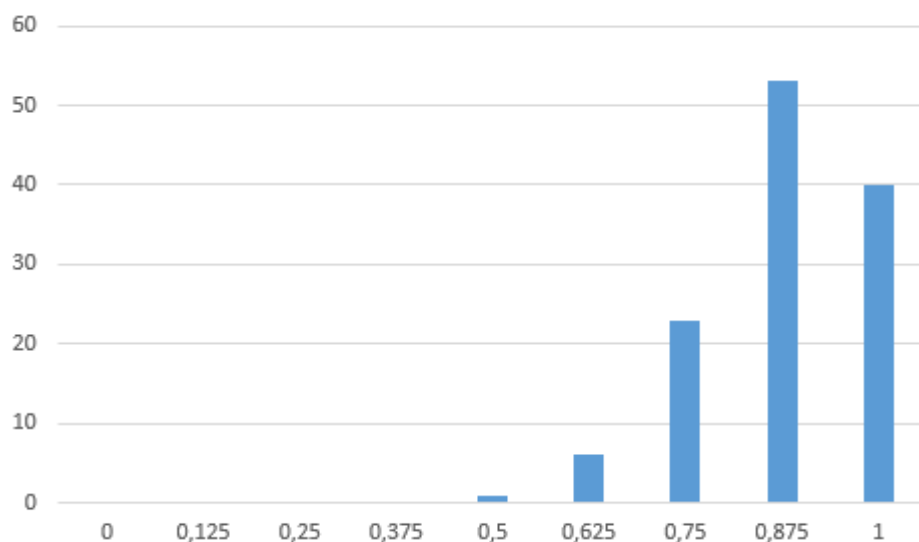


Рисунок 1—распределение GC-контента потерянных слов длины 8 хромосомы AJ270058 (вид *Arabidopsis thaliana*).

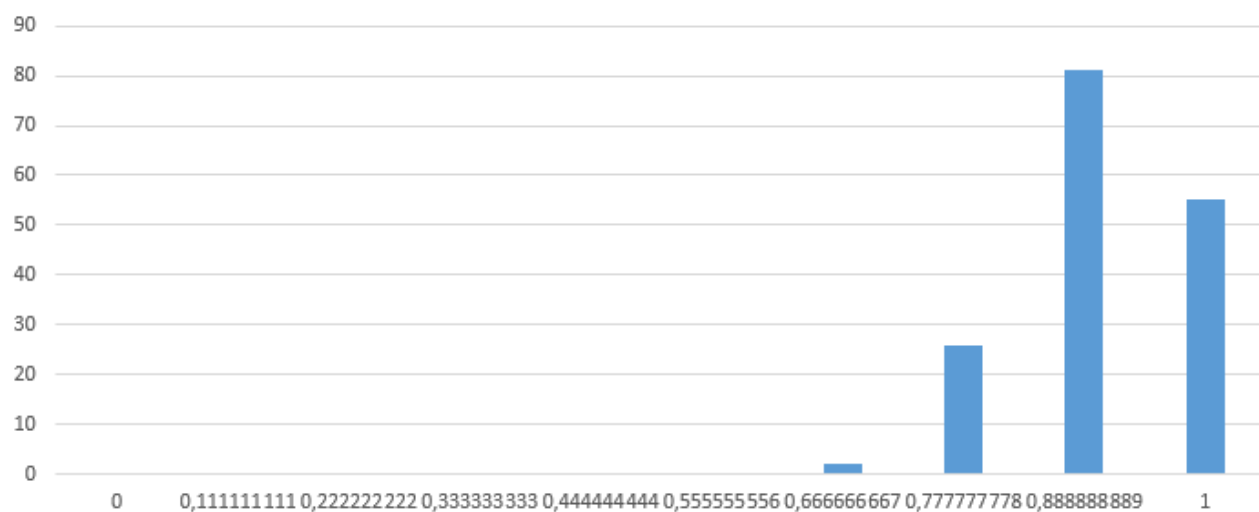


Рисунок 2 – распределение GC-контента потерянных слов длины 9 хромосомы BA000014 (вид *Arabidopsis thaliana*).

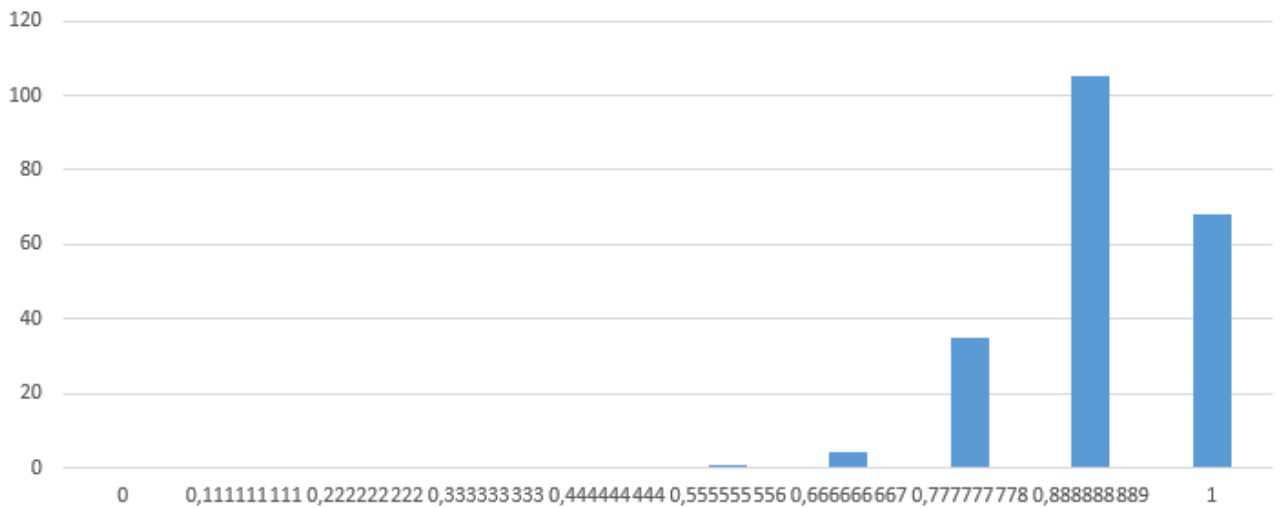


Рисунок 3 – распределение GC-контента потерянных слов длины 9 хромосомы BA000015 (вид *Arabidopsis thaliana*).

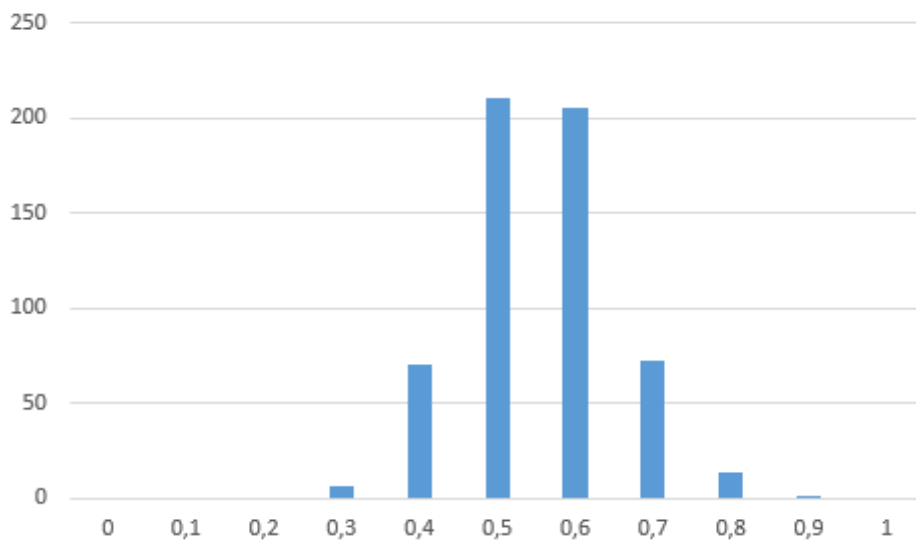


Рисунок 4 – распределение GC-контента потерянных слов длины 10 хромосомы cm000360 (вид *Anopheles gambiae*).

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Кутыркин, В.А. Распознавание скрытой периодичности в последовательностях ДНК / В.А. Кутыркин, М.Б. Чалей // Математическая биология и биоинформатика. – 2013. – Т. 8, № 2. – С. 502–512.
2. Мирошниченко, Л.А. Программа «SIM_PAT» для поиска фракталоподобных структур в ДНК-последовательностях / Л.А.Мирошниченко, В.Д. Гусев // Хроники объединенного фонда электронных ресурсов наука и образование – 2014. – № 1 (56). – С. 62.
3. Спангенберг, В.Е. Повторяющиеся последовательности ДНК могут принимать участие в формировании синаптомных комплексов / В.Е. Спангенберг, С.Я. Дадашев, О.Л. Коломиец, Ю.Ф. Богданов // Цитология. – М., 2011. – Т. 53, № 3. – С. 235–241.
4. Rawat, A. Novel implementation of conditional co-regulation by graph theory to derive co-expressed genes from microarray data / A. Rawat, Y. Deng // BioMed Central Bioinformatics. – Oklahoma City, OK, USA, 2008.
5. Hao, B. Avoided strings in bacterial complete genomes and a related combinatorial problem / B. Hao, H. Xie, Z. Yu, G. Chen // Annals of combinatorics. – 1998(2000). – #4 – P. 247–255.
6. Sadovsky, M. Lost Strings in Genomes: What Sense Do They Make? / M. Sadovsky et al. // Strong Inhomogeneity in Triplet Distribution. – P. 20-29.
7. Сайт рестрикции бактерии *Acetobacter pasteurianus* [Электронный ресурс] – режим доступа:
http://www.skygen.com/catalog/biohimicheskie_reaktivy/new_england_biolabs/endo_nukleazy_restriksii/standartnye_restriktazy/apali/

ПРИЛОЖЕНИЕ А

```
#include <stdio.h>
#include <tchar.h>
#include <string.h>
#include <math.h>
#include <stdlib.h>
#include <windows.h>
#include <stdlib.h>
char filepath[100];
char ifilepath[100], ofilepath[100], PathOfMatch[100], lenghtResult[150];
char pathoftables[100];
const int maxfrnumber=1500000000, froverlap=100, maxwordsnumber=2000000;
int q=1, wc=0, fwc=0, orq, i, j, k, NFile, NWords, i1, i2, frnumber, wordsnumber;
int *wordcounter, wn, nword[20], j11, j2, j9, *filecounter, spaceforfiles=0;
int spaceforwords;
char ifilename[150], ofilename[150], FoNW[150];
FILE *f, *f1, *f2, *f3, *f4;
char
fout2[91], string[10], *fout, **qout, yorn, mode[10], *lslash, *dot, fout3[150], fullpath
[150], fullpath2[150], copy[20], **arrayofwords, **arrayoffiles, *semicolon, *copy1;
bool sq=false, *wordinside, serror, pmfon=false, **foundfiles;
double rwn;
WIN32_FIND_DATA data2;
HANDLE hFind;
void fmain();
void setq();
void pm_file();
void packetmode();
void match();
void lenght_split();
void lenght_match();
void table_match();
int main(int argc, char* argv[])
{
    serror=false;
    if (argc==1)
    {
        printf("error");
        return 2;
    }
    else
```

```

{
    for (i=1;i<argc;i++)
    {
        if (strcmp(argv[i],"-m")==0) strcpy(mode,argv[i+1]);
        if (strcmp(mode,"of")==0)
        {
            if (strcmp(argv[i],"-ifn")==0)
strcpy(ifilename,argv[i+1]);
            if (strcmp(argv[i],"-ofn")==0)
strcpy(ofilename,argv[i+1]);
            if (strcmp(argv[i],"-fp")==0)
strcpy(filepath,argv[i+1]);
            if (strcmp(argv[i],"-orw")==0)
            {
                strcpy(copy,argv[i+1]);
                orq=atoi(copy);
            };
        };
        if(strncmp(mode,"pm",2)==0)
        {
            if (strcmp(argv[i],"-fp")==0)
strcpy(filepath,argv[i+1]);
            if (strcmp(argv[i],"-ofp")==0)
strcpy(ofilepath,argv[i+1]);
            if (strcmp(argv[i],"-ifp")==0)
strcpy(ifilepath,argv[i+1]);
            if (strncmp(mode,"pmm",3)==0)
            {
                if (strcmp(argv[i],"-pom")==0)
strcpy(PathOfMatch,argv[i+1]);
                if (strncmp(mode,"pmmt",4)==0)
                {
                    if (strcmp(argv[i],"-t")==0)
strcpy(pathoftables,argv[i+1]);
                    if (strncmp(mode,"pmmt1",5)==0)
                    {
                        if (strcmp(argv[i],"-lr")==0)
strcpy(lenghtResult,argv[i+1]);
                    };
                };
            };
            if (strncmp(mode,"pmml",4)==0)
            {

```

```

                                if (strcmp(argv[i], "-lr")==0)
strcpy(lenghtResult, argv[i+1]);
                                };
                                };
                                };
                                if (strncmp(mode, "m", 1)==0)
                                {
                                    if (strcmp(argv[i], "-ifp")==0)
strcpy(ifilepath, argv[i+1]);
                                    if (strcmp(argv[i], "-fp")==0)
strcpy(filepath, argv[i+1]);
                                    if (strcmp(argv[i], "-pom")==0)
strcpy(PathOfMatch, argv[i+1]);
                                    if (strncmp(mode, "mt", 2)==0)
                                    {
                                        if (strcmp(argv[i], "-t")==0)
strcpy(pathoftables, argv[i+1]);
                                        if (strncmp(mode, "mtl", 3)==0)
                                        {
                                            if (strcmp(argv[i], "-lr")==0)
strcpy(lenghtResult, argv[i+1]);
                                            };
                                        };
                                        if (strncmp(mode, "ml", 2)==0)
                                        {
                                            if (strcmp(argv[i], "-lr")==0)
strcpy(lenghtResult, argv[i+1]);
                                            };
                                        };
                                    };
                                };
                                };
                                if (strcmp(mode, "of")==0)
                                {
                                    fmain();
                                };
                                if (strncmp(mode, "pm", 2)==0)
                                {
                                    packetmode();
                                    if (strncmp(mode, "pmm", 3)==0)
                                    {
                                        strcpy(ifilepath, ofilepath);
                                        strcat(ifilepath, "*");
                                        match();
                                    }
                                }

```

```

        if (strncmp (mode, "pmmt", 4) == 0)
        {
            table_match ();
            if (strncmp (mode, "pmmt1", 5) == 0)
                lenght_match ();
        };
        if (strncmp (mode, "pmm1", 4) == 0) lenght_match ();
    };
};
if (strncmp (mode, "m", 1) == 0)
{
    match ();
    if (strncmp (mode, "mt", 2) == 0)
    {
        table_match ();
        if (strncmp (mode, "mt1", 3) == 0)
            lenght_match ();
    };
    if (strncmp (mode, "m1", 2) == 0) lenght_match ();
};
if (serror)
    return 1;
return 0;
}
void fmain ()
{
    rwn=pow(4.0, orq);
    if (rwn<maxwordsnumber)
    {
        wordsnumber=rwn;
    }
    else wordsnumber = maxwordsnumber;
    qout = new char*[wordsnumber];
    for (i=0; i<wordsnumber; i++) qout[i]=new char[orq+1];
    wordinside = new bool[wordsnumber];
    f = fopen(ifilename, "r");
    if (!f)
    {
        printf("can't open reading file!");
        serror=true;
        goto end_fmain;
    };
    strcpy(fullpath, filepath);

```



```

strcat(fullpath, "dna.txt");
f1 = fopen(fullpath, "w");
if (!f1)
{
    printf("can't open file to write!");
    serror=true;
    goto end_fmain;
};
if(!pmfon)
{
    f2 = fopen(ofilename, "w");
    if (!f2)
    {
        printf("can't open file of output!");
        serror=true;
        goto end_fmain;
    };
    fclose(f2);
};
while (!feof(f))
{
    fgets(fout2, 91, f);
    if (strncmp(fout2, "SQ", 2) == 0)
    {
        sscanf(fout2,"%*s %*s %s",copy);
        frnumber = atoi(copy);
        sq = true;
        continue;
    };
    if (sq && (strlen(fout2)>5))
    {
        for (i=0;i<6;i++)
        {
            strncpy(string, fout2 + 5 + i * 11, 10);
            fprintf(f1, "%10.10s", string);
        };
    };
};
fclose(f);
fclose(f1);
if (frnumber>maxfrnumber) frnumber=maxfrnumber;
fout=new char[frnumber+1];
strcpy(fullpath, filepath);

```

```

strcat(fullpath, "dna.txt");
f=fopen(fullpath, "r");
fgets(fout, frnumber+1, f);
strcpy(fullpath, filepath);
j = 1;
sprintf(fout2, "NF%03d.txt", j);
strcat(fullpath, fout2);
f1 = fopen(fullpath, "w");
fprintf(f1, "%s", fout);
if ((!feof(f)) && (fgetc(f) != ' '))
{
    do
    {
        fclose(f1);
        fclose(f);
        j++;
        strcpy(fullpath, filepath);
        strcat(fullpath, "dna.txt");
        f = fopen(fullpath, "r");
        strcpy(fullpath, filepath);
        sprintf(fout2, "NF%03d.txt", j);
        strcat(fullpath, fout2);
        f1 = fopen(fullpath, "w");
        for (i=0; i<j-1; i++) fgets(fout, frnumber-froverlap+1, f);
        fgets(fout, frnumber+1, f);
        fprintf(f1, "%s", fout);
    }while (!feof(f));
}
else j++;
fclose(f1);
fclose(f);
NFile = j;
while (q <= orq)
{
    setq();
    q++;
};
q--;
strcpy(fullpath, filepath);
strcat(fullpath, "qnext.txt");
f1 = fopen(fullpath, "r");
j = 1;
while (!feof(f1))
{

```

```

strcpy(fullpath, filepath);
sprintf(fout2, "NW%03d.txt", j);
strcat(fullpath, fout2);
f3 = fopen(fullpath, "w");
strcpy(fullpath, filepath);
sprintf(fout2, "W%03d.txt", j);
strcat(fullpath, fout2);
f = fopen(fullpath, "w");
for (i = 0; i<wordsnumber; i++)
{
    fscanf(f1, "%s ", copy);
    fprintf(f3, "0 ");
    fprintf(f, "%s ", copy);
    if (feof(f1))break;
};
fclose(f);
fclose(f3);
j++;
};
fclose(f1);
NWords=j;
if(!pmfon)
{
    strcpy(fullpath, filepath);
    strcat(fullpath, "nmw.txt");
    f2=fopen(fullpath, "w");
};
for (i = 1; i<NFile; i++)
{
    fwc = 0;
    strcpy(fullpath, filepath);
    sprintf(fout2, "NF%03d.txt", i);
    strcat(fullpath, fout2);
    f = fopen(fullpath, "r");
    fscanf(f, "%s", fout);
    fclose(f);
    for (j = 1; j<NWords; j++)
    {
        strcpy(fullpath, filepath);
        sprintf(fout2, "NW%03d.txt", j);
        strcat(fullpath, fout2);
        f3 = fopen(fullpath, "r");
        strcpy(fullpath, filepath);

```

```

sprintf(fout2, "W%03d.txt", j);
strcat(fullpath, fout2);
f1 = fopen(fullpath, "r");
for (k = 0; k<wordsnumber; k++)
{
    fscanf(f1, "%s ", qout[k]);
    fscanf(f3, "%s ", copy);
    i1 = atoi(copy);
    if (i1 == 0)
    {
        wordinside[k] = false;
    }
    else wordinside[k] = true;
    if (feof(f1))break;
};
i2 = k;
fclose(f1);
fclose(f3);
strcpy(fullpath, filepath);
sprintf(fout2, "NW%03d.txt", j);
strcat(fullpath, fout2);
f3 = fopen(fullpath, "w");
for (k = 0; k<=i2; k++)
{
    if (!wordinside[k])
    {
        if (strstr(fout, qout[k]) == NULL)
        {
            if (pmfon)
            {
                fclose(f3);
                strcpy(fullpath, filepath);
                strcat(fullpath, "fonw.txt");
                f4=fopen(fullpath,"a");
                fprintf(f4,"%s %d\n",
data2.cFileName,org);

                fclose(f4);
                pmfon=false;
                goto end_fmain_m;
            };
            fprintf(f3, "0 ");
            fwc++;
        }
    }
}

```

```

        else fprintf(f3, "1 ");
    }
    else fprintf(f3, "1 ");
    if ((!pmfon)&&((k+1)%5000==0)) printf("number of words=
%d, number of missing words= %d\n",k+1, fwc);
    };
    fclose(f3);
};
if(!pmfon) fprintf(f2, "%d %d\n", i, fwc);
};
if(!pmfon)
{
    fclose(f2);
    f2 = fopen(ofilename, "w");
    if (!f2) { printf("can't open file!");  serror=true; goto
end_fmain;};
    fwc = 0;
    for (j=1;j<NWords;j++)
    {
        strcpy(fullpath,filepath);
        sprintf(fout2,"W%03d.txt",j);
        strcat(fullpath,fout2);
        f1=fopen(fullpath,"r");
        for (k=0;k<wordsnumber;k++)
            fscanf(f1, "%s ", qout[k]);
        fclose(f1);
        strcpy(fullpath,filepath);
        sprintf(fout2,"NW%03d.txt",j);
        strcat(fullpath,fout2);
        f3=fopen(fullpath,"r");
        for (k=0;k<wordsnumber;k++)
        {
            fscanf(f3,"%s ",copy);
            i1=atoi(copy);
            if (i1==0)
            {
                fwc++;
                fprintf(f2,"%s ",qout[k]);
            };
        };
        fclose(f3);
    };
    fclose(f2);
};

```

```

};
end_fmain_m:
    delete [] fout;
    delete [] wordinside;
    for (k=0;k<wordsnumber;k++)
        delete [] qout[k];
    delete [] qout;
end_fmain;;
}
void setq()
{
    FILE *f, *f1;
    char finput[20],foutput[20],v;
    if (q == 1)
    {
        strcpy(fullpath, filepath);
        strcat(fullpath, "qnext.txt");
        f = fopen(fullpath, "w");
        fputs("a c g t ", f);
        fclose(f);
    }
    else
    {
        strcpy(fullpath, filepath);
        strcat(fullpath, "qprev.txt");
        f = fopen(fullpath, "r");
        strcpy(fullpath, filepath);
        strcat(fullpath, "qnext.txt");
        f1 = fopen(fullpath, "w");
        while (!feof(f))
        {
            fscanf(f, "%s ", foutput);
            strcpy(finput, foutput);
            strcat(finput, "a ");
            fputs(finput, f1);
            strcpy(finput, foutput);
            strcat(finput, "c ");
            fputs(finput, f1);
            strcpy(finput, foutput);
            strcat(finput, "g ");
            fputs(finput, f1);
            strcpy(finput, foutput);
            strcat(finput, "t ");

```

```

        fputs(fininput, f1);
    };
    fclose(f);
    fclose(f1);
};
strcpy(fullpath, filepath);
strcat(fullpath, "qprev.txt");
f = fopen(fullpath, "w");
strcpy(fullpath, filepath);
strcat(fullpath, "qnext.txt");
f1 = fopen(fullpath, "r");
while (!feof(f1))
{
    v = fgetc(f1);
    if (v>0) fputc(v, f);
};
fclose(f);
fclose(f1);
}
void packetmode ()
{
    pm_file();
    strcpy(FoNW, filepath);
    strcat(FoNW, "fonw.txt");
    hFind=FindFirstFileA(ifilepath, &data2);
    if (hFind != INVALID_HANDLE_VALUE)
    {
        if( data2.dwFileAttributes != FILE_ATTRIBUTE_DIRECTORY )
        {
            if (strcmp(ofilepath, "\\0")==0)
            {
                printf("Can't open output directory!");
                serror=true;
                goto end_pm;
            };
            f4=fopen(FoNW, "r");
            if (!f4)
            {
                printf("Can't open file of numbers of symbols of
possibly missed words!");
                serror=true;
                goto end_pm;
            };
        }
    }
}

```

```

    strncpy(ifilename, ifilepath, strlen(ifilepath)-1);
    ifilename[strlen(ifilepath)-1]='\0';
    strcat(ifilename, data2.cFileName);
    lslash=strrchr(ifilename, '\\');
    lslash++;
    dot=strchr(lslash, '.');
    strcpy(ofilename, ofilepath);
    strcat(ofilename, lslash, dot-lslash);
    strcat(ofilename, ".avs");
    while (!feof(f4))
    {
        fgets(fout3, 150, f4);
        if (strncmp(lslash, fout3, strlen(lslash)) == 0)
        {
            sq=true; sscanf(fout3, "%*s
%s", copy); orq=atoi(copy); break;
        };
    };
    fclose(f4);
    if (!sq)
    {
        printf("Can't find number of symbols of possibly missed
words in file. If you DOESN'T want to continue write n ");
        scanf("%c", yorn);
        if (yorn=='n') {goto end_pm;}
        else goto continue_pm;
    };
    sq=false; q=1; fwc=0;
    fmain();
}
else printf("directory\n");
continue_pm:
do
{
    if( data2.dwFileAttributes != FILE_ATTRIBUTE_DIRECTORY )
    {
        if (strcmp(ofilepath, "\\0")==0)
        {
            printf("Can't open output directory!");
            serror=true;
            goto end_pm;
        };
        f4=fopen(FoNW, "r");

```



```

        if (!f4)
        {
            printf("Can't open file of numbers of symbols of
possibly missed words!");
            serror=true;
            goto end_pm;
        };
        strncpy(ifilename,ifilepath,strlen(ifilepath)-1);
        ifilename[strlen(ifilepath)-1]='\0';
        strcat(ifilename,data2.cFileName);
        lslash=strrchr(ifilename,'\\');
        lslash++;
        dot=strchr(lslash, '.');
        strcpy(ofilename,ofilepath);
        strncat(ofilename,lslash,dot-lslash);
        strcat(ofilename, ".avs");
        while (!feof(f4))
        {
            fgets(fout3, 150, f4);
            if (strncmp(lslash, fout3, strlen(lslash)) == 0)
            {
                sq = true;
                sscanf(fout3,"%*s %s",copy);
                orq = atoi(copy);
                break;
            };
        };
        fclose(f4);
        if (!sq)
        {
            printf("Can't find number of symbols of possibly
missed words in file. If you DOESN'T want to continue write n ");
            scanf("%c",yorn);
            if (yorn=='n') {goto end_pm;}
            else continue;
        };
        sq=false;fwc=0;q=1;
        printf("working with file %s, number of symbols
%d\n",ifilename,orq);
        fmain();
    }
    else printf("directory\n");
}while(FindNextFileA(hFind,&data2));

```

```

        }else printf("Can't open input directory!");
end_pm;;
}
void pm_file()
{
    hFind=FindFirstFileA(ifilepath,&data2);
    if (hFind != INVALID_HANDLE_VALUE)
    {
        strcpy(fullpath,filepath);
        strcat(fullpath, "fonw.txt");
        f4=fopen(fullpath,"w");
        fclose(f4);
        pmfon=true;
        if( data2.dwFileAttributes != FILE_ATTRIBUTE_DIRECTORY )
        {
            strncpy(ifilename,ifilepath,strlen(ifilepath)-1);
            ifilename[strlen(ifilepath)-1]='\0';
            strcat(ifilename,data2.cFileName);
            for (orq=1;orq<15;orq++)
            {
                sq=false;q=1;fwc=0;
                fmain();
                if (!pmfon)
                {
                    pmfon=true;break;
                };
            };
        }
        else printf("directory\n");
        do
        {
            if( data2.dwFileAttributes != FILE_ATTRIBUTE_DIRECTORY )
            {
                strncpy(ifilename,ifilepath,strlen(ifilepath)-1);
                ifilename[strlen(ifilepath)-1]='\0';
                strcat(ifilename,data2.cFileName);
                for (orq=1;orq<15;orq++)
                {
                    sq=false;q=1;fwc=0;
                    fmain();
                    if (!pmfon)
                    {
                        pmfon=true;break;

```

```

};
};
printf("worked with file %s, calculated number of
symbols %d\n",ifilename,orq);
}
else printf("directory\n");
}while(FindNextFileA(hFind,&data2));
pmfon=false;
}else printf("Can't open input directory!");
}
void match()
{
strcpy(fullpath,filepath);
strcat(fullpath,"fonw.txt");
f2=fopen(fullpath,"r");
if (!f2)
{
printf("Can't open file of numbers of symbols of possibly missed
words!");
serror=true;
goto end_match;
};
fclose(f2);
strcpy(fullpath,filepath);
strcat(fullpath,"res_filenames.txt");
f=fopen(fullpath,"w");
fclose(f);
filecounter=new int [20];
for (i=0;i<20;i++)
filecounter[i]=0;
lenght_split();
for (i=0; i<20; i++)
if (spaceforfiles<((filecounter[i]*10)+1))
spaceforfiles=(filecounter[i]*10)+1;
delete [] filecounter;
for (orq=1;orq<20;orq++)
{
printf("\nq=%d\n",q);
sprintf(copy,"%d",orq);
strcpy(fullpath,filepath);
strcat(fullpath,"fonw");
strcat(fullpath,copy);
strcat(fullpath,".txt");

```

```

f3=fopen(fullpath,"r");
if (!f3) continue;
fclose(f3);
strncpy(fullpath,ifilepath,strlen(ifilepath)-2);
fullpath[strlen(ifilepath)-2]='\0';
lslash=strrchr(fullpath,'\\');
lslash++;
dot=strchr(fullpath,'\0');
strcpy(fullpath2,PathOfMatch);
strncat(fullpath2,lslash,dot-lslash);
strcat(fullpath2, copy);
strcat(fullpath2, ".txt");
f=fopen(fullpath2,"w");
if (!f)
{
    printf("can't open file to write!");
    serror=true;
    goto end_match;
};
fclose(f);
strcpy(fullpath,filepath);
strcat(fullpath, "res_filenames.txt");
f=fopen(fullpath,"a");
fprintf(f,"%s %d\n",fullpath2,orq);
fclose(f);
rwn=pow(4.0,orq);
if (rwn<maxwordsnumber)
{wordsnumber=rwn;}
else wordsnumber = maxwordsnumber;
qout=new char*[wordsnumber];
wordcounter=new int[wordsnumber];
arrayoffiles=new char*[wordsnumber];
for (i=0; i<wordsnumber; i++)
{
    qout[i]=new char[orq+1];
    arrayoffiles[i]=new char[spaceforfiles];
    strcpy(arrayoffiles[i]," ");
    wordcounter[i]=0;
};
q=1;
while (q <= orq)
{
    setq();

```

```

        q++;
    };
    q--;
    strcpy(fullpath, filepath);
    strcat(fullpath, "qnext.txt");
    f1=fopen(fullpath, "r");
    j=1;
    while (!feof(f1))
    {
        strcpy(fullpath, filepath);
        sprintf(fout2, "W%03d.txt", j);
        strcat(fullpath, fout2);
        f = fopen(fullpath, "w");
        for (i = 0; i<wordsnumber; i++)
        {
            fscanf(f1, "%s ", copy);
            fprintf(f, "%s ", copy);
            if (feof(f1))break;
        };
        fclose(f);
        j++;
    };
    fclose(f1);
    NWords = j;
    copy[0]='\0';
    hFind=FindFirstFileA(ifilepath, &data2);
    if (hFind != INVALID_HANDLE_VALUE)
    {
        if( data2.dwFileAttributes != FILE_ATTRIBUTE_DIRECTORY )
        {
            strncpy(ifilename, ifilepath, strlen(ifilepath)-1);
            ifilename[strlen(ifilepath)-1]='\0';
            strcat(ifilename, data2.cFileName);
            lslash=strrchr(ifilename, '\\');
            lslash++;
            dot=strchr(lslash, '.');
            sprintf(copy, "%d", orq);
            strcpy(fullpath, filepath);
            strcat(fullpath, "fonw");
            strcat(fullpath, copy);
            strcat(fullpath, ".txt");
            f3=fopen(fullpath, "r");
            while (!feof(f3))

```



```

        for (i1=0;i1<k;i1++)
        {
            for (i=0; i<wordnumber; i++)
                if
(strcmp(qout[i],arrayofwords[i1])==0)
                {

                    strncat(arrayoffiles[i],lslash,dot-lslash);

                                strcat(arrayoffiles[i],";");
                                wordcounter[i]++;
                                break;

                                };

                };

        };
        for (i=0;i<k;i++)delete [] arrayofwords[i];
    }
else printf("directory\n");
continue_match:
do
{
    if( data2.dwFileAttributes!= FILE_ATTRIBUTE_DIRECTORY)
    {

        strncpy(ifilename,ifilepath,strlen(ifilepath)-1);
        ifilename[strlen(ifilepath)-1]='\0';
        strcat(ifilename,data2.cFileName);
        lslash=strrchr(ifilename,'\\');
        lslash++;
        dot=strchr(lslash,'. ');
        sprintf(copy,"%d",orq);
        strcpy(fullpath,filepath);
        strcat(fullpath, "fonw");
        strcat(fullpath, copy);
        strcat(fullpath, ".txt");
        f3=fopen(fullpath,"r");
        while (!feof(f3))
        {

            fgets(fout3, 150, f3);
            sq=true;
            if (strncmp(lslash,fout3,strlen(lslash)-
3)==0)

            {

                sq=false;
                break;

```

```

};
};
fclose(f3);
copy[0]='\0';
if(sq) continue;
k=0;sq=false;
printf("working with file %s\n",ifilename);
f=fopen(ifilename,"r");
while (!feof(f))
{
    fscanf(f, "%s ",copy);
    k++;
    if (strcmp(copy,"\0")==0)
        sq=true;
};
fclose(f);
if(sq) continue;
f=fopen(ifilename,"r");
arrayofwords=new char*[k];
for (i=0; i<k; i++)
{
    fscanf(f, "%s ",copy);
    arrayofwords[i]=new char[orq+1];
    strcpy(arrayofwords[i],copy);
};
fclose(f);
for (j = 1; j<NWords; j++)
{
    strcpy(fullpath, filepath);
    sprintf(fout2, "W%03d.txt", j);
    strcat(fullpath, fout2);
    f1 = fopen(fullpath, "r");
    for (i=0; i<wordsnumber; i++)
    {
        fscanf(f1, "%s ", qout[i]);
    };
    fclose(f1);
    for (i1=0;i1<k;i1++)
    {
        for (i=0; i<wordsnumber; i++)
            if
(strcmp(qout[i],arrayofwords[i1])==0)
                {

```



```

strncat(arrayoffiles[i],lslash,dot-lslash);

strcat(arrayoffiles[i],";");

wordcounter[i]++;
break;

};

};

for (i=0;i<k;i++)delete [] arrayofwords[i];
delete [] arrayofwords;

}

else printf("directory\n");
}while(FindNextFileA(hFind,&data2));
f=fopen(fullpath2,"w");
for (i=0; i<wordsnumber; i++)
    if (wordcounter[i]!=0)
        fprintf(f,"%s
%d%s\n",qout[i],wordcounter[i],arrayoffiles[i]);
fclose(f);
}else printf("Can't open input directory!");
delete [] wordcounter;
for (k=0;k<wordsnumber;k++)
{
    delete [] qout[k];
    delete [] arrayoffiles[k];
};
delete [] qout;
delete [] arrayoffiles;

};

end_match;;
printf("end of match");
}

void lenght_split()
{
    strcpy(fullpath,filepath);
    strcat(fullpath, "fonw.txt");
    f2=fopen(fullpath,"r");
    while (!feof(f2))
    {
        fgets(fout3, 150, f2);
        sscanf(fout3,"%*s %s",copy);
        strcpy(fullpath,filepath);

```

```

        strcat(fullpath, "fonw");
        strcat(fullpath, copy);
        strcat(fullpath, ".txt");
        f3=fopen(fullpath,"w");
        fclose(f3);
};
fclose(f2);
strcpy(fullpath, filepath);
strcat(fullpath, "fonw.txt");
f2=fopen(fullpath,"r");
while (!feof(f2))
{
    fgets(fout3, 150, f2);
    if (strcmp(fout3, "\0")==0) break;
    sscanf(fout3, "%*s %s", copy);
    i=atoi(copy);
    filecounter[i-1]++;
    strcpy(fullpath, filepath);
    strcat(fullpath, "fonw");
    strcat(fullpath, copy);
    strcat(fullpath, ".txt");
    f3=fopen(fullpath,"a");
    fputs(fout3, f3);
    fclose(f3);
};
fclose(f2);
}
void lenght_match()
{
    printf("starting lenght_match...");
    arrayofwords=new char*[1];
    arrayofwords[0]=new char[1000];
    strcpy(arrayofwords[0], "\0");
    for (i=0; i<20; i++) nword[i]=0;
    strcpy(fullpath, filepath);
    strcat(fullpath, "res_filenames.txt");
    f=fopen(fullpath, "r");
    while (!feof(f))
    {
        fscanf(f, "%s %s\n", fullpath2, copy);
        orq=atoi(copy);
        f1=fopen(fullpath2, "r");
        while (!feof(f1))

```



```

        i1=i+nword[orq-2];
        i2=i1+nword[orq-1];
        for (j=i;j<i1;j++)
        {
            sscanf(arrayofwords[j],"%s %*s %*s",copy);
            for (k=i1;k<i2;k++)
                if (strstr(arrayofwords[k],copy)!=NULL)
                {

fprintf(f,"%s%s\n",arrayofwords[k],arrayofwords[j]);

                };

        };

    };
fclose(f);
printf("done\n");
}
void table_match()
{
    spaceforwords=spaceforfiles+26;
    copy1=new char[spaceforwords];
    printf("starting table_match...");
    arrayofwords=new char*[1];
    arrayofwords[0]=new char[spaceforwords];
    strcpy(arrayofwords[0],"\\0");
    for (i=0;i<20;i++) nword[i]=0;
    strcpy(fullpath,filepath);
    strcat(fullpath, "res_filenames.txt");
    f=fopen(fullpath,"r");
    while(!feof(f))
    {
        fscanf(f,"%s %s\n",fullpath2,copy);
        orq=atoi(copy);
        f1=fopen(fullpath2,"r");
        while(!feof(f1))
        {
            fgets(arrayofwords[0],spaceforwords,f1);
            if (strcmp(arrayofwords[0],"\\0")!=0) nword[orq-1]++;
            strcpy(arrayofwords[0],"\\0");
        };
        fclose(f1);
    };
fclose(f);
delete [] arrayofwords[0];

```

```

delete [] arrayofwords;
NWords=0;
for (i=0;i<20;i++) NWords+=nword[i];
arrayofwords=new char*[NWords+1];
for (i=0; i<NWords+1; i++)
{
    arrayofwords[i]=new char[spaceforwords];
    strcpy(arrayofwords[i],"\0");
};
i=0;
f=fopen(fullpath,"r");
while(!feof(f))
{
    fscanf(f,"%s %s\n",fullpath2,copy);
    orq=atoi(copy);
    f1=fopen(fullpath2,"r");
    while(!feof(f1))
    {
        fgets(arrayofwords[i],spaceforwords,f1);
        i++;
    };
    fclose(f1);
    if (i<NWords)
    {
        if (strcmp(arrayofwords[i],"\0")==0) i--;
    }
    else i--;
};
fclose(f);
for (orq=20;orq>0;orq--)
    if (nword[orq-1]!=0)
    {
        i=0;
        for (j=0;j<orq-1;j++) i+=nword[j];
        i1=i+nword[orq-1];
        arrayoffiles=new char*[i1-i];
        for (j=0;j<i1-i;j++)arrayoffiles[j]=new char[100];
        k=0;
        for (j=i;j<i1;j++)
        {
            sscanf(arrayofwords[j],"%*s %*s %s",copy1);
            semicolon=strchr(copy1,';');
            strncpy(arrayoffiles[k],copy1,semicolon-copy1);

```

```

arrayoffiles[k][semicolon-copy1]='\0';
for (i2=0;i2<k;i2++)
    if (strcmp(arrayoffiles[k],arrayoffiles[i2])==0)
    {
        k--;
        break;
    };
k++;
strcpy(copy1,semicolon+1);
semicolon=strchr(copy1,';');
while (semicolon!=NULL)
{
    strncpy(arrayoffiles[k],copy1,semicolon-copy1);
    arrayoffiles[k][semicolon-copy1]='\0';
    for (i2=0;i2<k;i2++)

if (strcmp(arrayoffiles[k],arrayoffiles[i2])==0)
    {
        k--;
        break;
    };
k++;
semicolon=strchr(copy1,';');
if (semicolon==NULL)
{
    k--;
    break;
};
strcpy(copy1,semicolon+1);
};
};
foundfiles=new bool*[i1-i];
for (j=0;j<i1-i;j++)
{
    foundfiles[j]=new bool[k];
    for (j2=0;j2<k;j2++)
        foundfiles[j][j2]=false;
};
for (j=i;j<i1;j++)
{
    sscanf(arrayofwords[j],"%*s %s %*s",copy);
    wn=atoi(copy);
    if (wn>1)

```

```

        {
            sscanf(arrayofwords[j], "%*s %*s %s", copy1);
            semicolon=strchr(copy1, ';');
            strncpy(copy, copy1, semicolon-copy1);
            copy[semicolon-copy1]='\0';
            for(j2=0;j2<k;j2++)
                if(strcmp(copy, arrayoffiles[j2])==0)
                    {
                        foundfiles[j-i][j2]=true;
                        strcpy(copy1, semicolon+1);
                        semicolon=strchr(copy1, ';');
                        if (semicolon!=NULL)
                            {
                                j2=-1;
                                strncpy(copy, copy1, semicolon-
copy1);
                                copy[semicolon-copy1]='\0';
                            };
                    };
        };
};
};
strncpy(fullpath, ifilepath, strlen(ifilepath)-2);
fullpath[strlen(ifilepath)-2]='\0';
lslash=strrchr(fullpath, '\\');
lslash++;
dot=strchr(fullpath, '\0');
sprintf(copy, "%d", orq);
strcpy(fullpath2, pathoftables);
strncat(fullpath2, lslash, dot-lslash);
strcat(fullpath2, copy);
strcat(fullpath2, ".txt");
f=fopen(fullpath2, "w");
for(j=0;j<20;j++)
{
    fputs(" ", f);
};
j11=0;
for(j=0;j<k;j++)
{
    fprintf(f, "%s ", arrayoffiles[j]);
    j11+=strlen(arrayoffiles[j])+1;
};
j11=j11/(2*k);

```

```

fputs("\n",f);
for (j=i;j<i1;j++)
{
    sscanf(arrayofwords[j],"%*s %s %*s",copy);
    wn=atoi(copy);
    if (wn>1)
    {
        sscanf(arrayofwords[j],"%s %*s %*s",copy);
        fprintf(f,"%s ",copy);
        j9=20-strlen(copy);
        for(i2=0;i2<j9;i2++) fputs(" ",f);
        for(j2=0;j2<k;j2++)
        {
            for(i2=0;i2<j11;i2++) fputs(" ",f);
            if(foundfiles[j-i][j2])
            {
                fputs("+",f);
            }
            else fputs("-",f);
            for(i2=0;i2<j11-1;i2++) fputs(" ",f);
        };
        fputs("\n",f);
    };
};
fclose(f);
for (j=0;j<i1-i;j++)
{
    delete [] arrayoffiles[j];
    delete [] foundfiles[j];
};
delete [] arrayoffiles;
delete [] foundfiles;
};
for (i=0; i<NWords; i++) delete [] arrayofwords[i];
delete [] arrayofwords;
delete [] copy1;
printf("done\n");
}

```