Описание работы программы

Перед началом процесса обработки программы следует **ОБАЗЯТЕЛЬНО** в произвольном порядке выполнить 4 действия:

1. Загрузить короткий файл
2. Загрузить длинный файл
3. Загрузить матрицу парных сравнений
4. Выбрать папку сохранения результатов

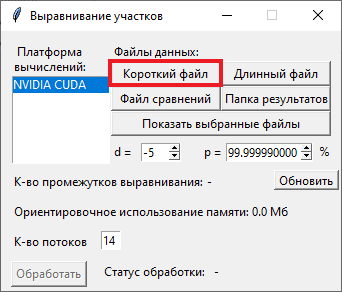
Также перед началом работы программы можно **В НЕОБЯЗАТЕЛЬНОМ ПОРЯДКЕ** выполнить 5 действия:

1. Задать платформу для параллельной обработки данных.
2. Задать штрафной коэффициент **d**
3. Задать отсечку **p** биномиального распределения
4. Задать количество потоков обработки
5. Просмотреть загруженные файлы

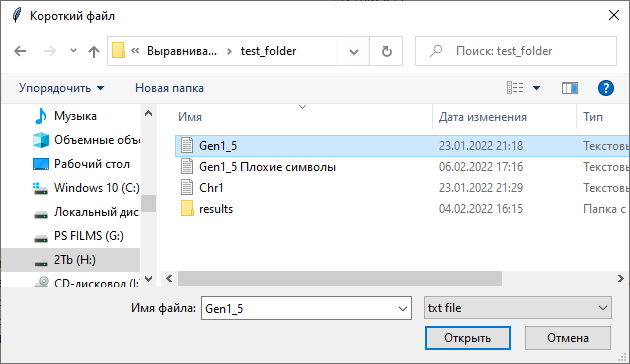
Обязательные действия

Загрузка короткого файла

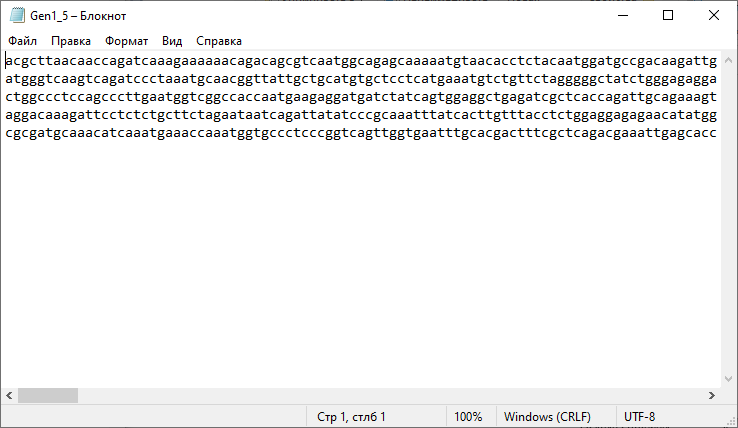
Для это надо нажать на кнопку главной формы «Короткий файл».



После чего указать в файловом диалоге путь до файла и нажать «Открыть». Допускаются только файлы с расширением txt.



Пример загруженного файла:



В момент загрузки в файле допускается наличие любых символов.

Символы пробела, табуляции, переноса строки и прочие не видимые символы удаляются из загруженных данных.

Все остальные символы по возможности приводятся к верхнему регистру. Таким образом стоит обратить внимание, что для программы нет различий между следующими строками:

* «aGCtTA»
* « AGCTTA»
* «agcctta»
* «AGC

TTA»

* «A

G

C

T

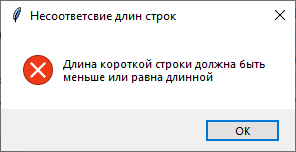
T

A»

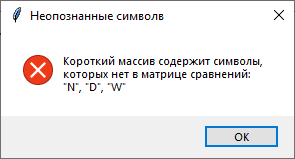
* «a gcc ta»
* «agc c ta»

Все эти строки приводятся к единому представлению: «AGCTTA»

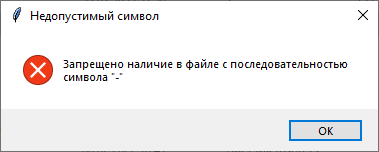
Типовые сообщения об ошибках:



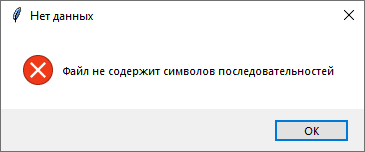
Возникает, когда длина только что загруженной короткой последовательности оказывается больше длины длинной последовательности. При появлении данной ошибки загрузка короткого файла прерывается и данные не сохраняются.



Возникает, если до этого был успешно загружен файл сравнений и среди символов в коротком файле встретились те, что не присутствуют в качестве элементов матрицы сравнения. При появлении данной ошибки загрузка короткого файла прерывается и данные не сохраняются.



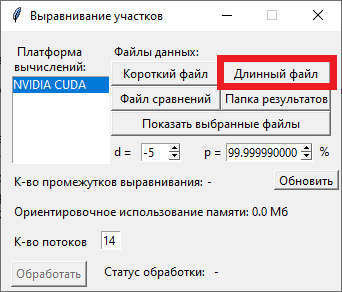
Возникает при обнаружении в файле с последовательностью символа «-», который в алгоритме выравнивания играет роль пропуска. При появлении данной ошибки загрузка длинного файла прерывается и данные не сохраняются.



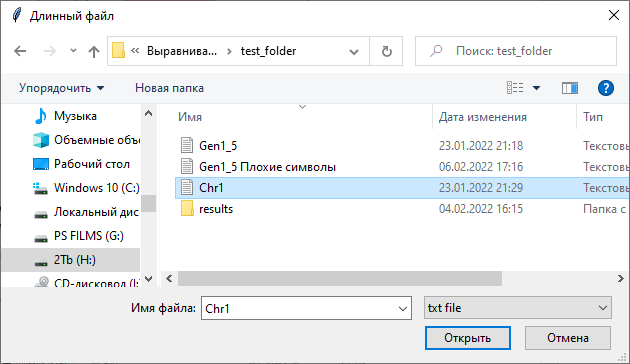
Возникает, когда при загрузке последовательности после удаление из загруженных данных всех пробелов, переносов строк и табуляций в последовательности не остаётся ни одного символа. При появлении данной ошибки загрузка длинного файла прерывается и данные не сохраняются.

Загрузка длинного файла

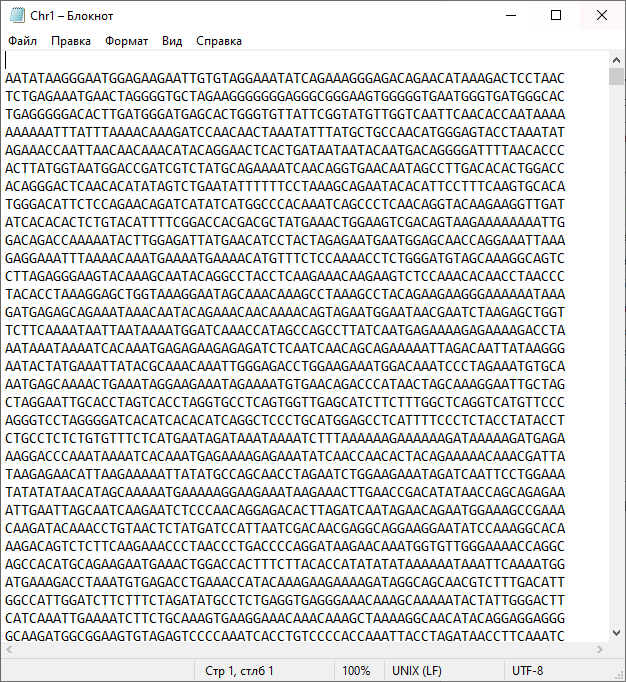
Для это надо нажать на кнопку главной формы «Длинный файл».



После чего указать в файловом диалоге путь до файла и нажать «Открыть». Допускаются только файлы с расширением txt.



Пример загруженного файла:



В момент загрузки в файле допускается наличие любых символов.

Символы пробела, табуляции, переноса строки и прочие не видимые символы удаляются из загруженных данных.

Все остальные символы по возможности приводятся к верхнему регистру. Таким образом стоит обратить внимание, что для программы нет различий между следующими строками:

* «aGCtTA»
* « AGCTTA»
* «agcctta»
* «AGC

TTA»

* «A

G

C

T

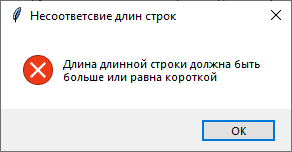
T

A»

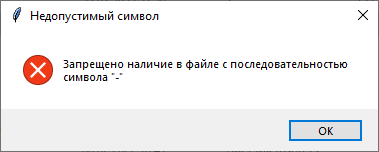
* «a gcc ta»
* «agc c ta»

Все эти строки приводятся к единому представлению: «AGCTTA»

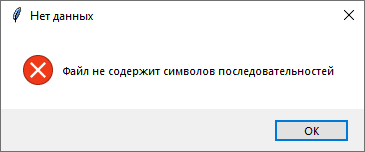
Типовые сообщения об ошибках:



Возникает, когда длина только что загруженной длинной последовательности оказывается меньше длины короткой последовательности. При появлении данной ошибки загрузка длинного файла прерывается и данные не сохраняются.



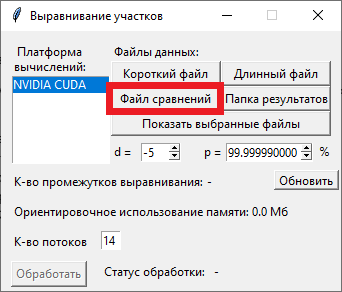
Возникает при обнаружении в файле с последовательностью символа «-», который в алгоритме выравнивания играет роль пропуска. При появлении данной ошибки загрузка длинного файла прерывается и данные не сохраняются.



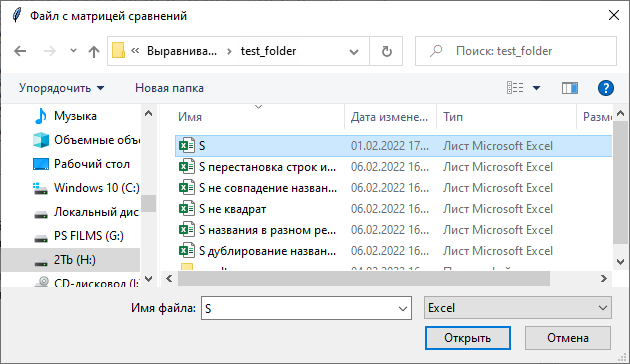
Возникает, когда при загрузке последовательности после удаление из загруженных данных всех пробелов, переносов строк и табуляций в последовательности не остаётся ни одного символа. При появлении данной ошибки загрузка длинного файла прерывается и данные не сохраняются.

Загрузка матрицы сравнений

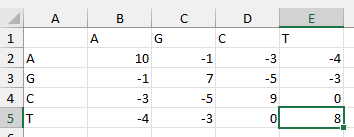
Для это надо нажать на кнопку главной формы «Файл сравнений».



После чего указать в файловом диалоге путь до файла и нажать «Открыть». Допускаются только файлы с расширением xlsx (Лист Microsoft Excel).



Пример правильно оформленного файла:



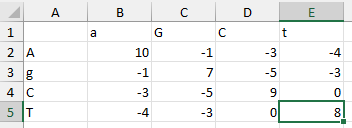
В качестве имен переменных допускаются:

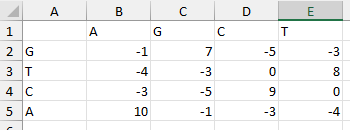
* Любые буквы
* Любые цифры (0-9)
* Пробел, табуляция, перенос строки
* «(», «=», «+», «%», «/», «:», «^», и т.д.

В качестве имен переменных **НЕ ДОПУСКАЮТСЯ**:

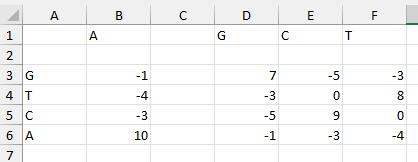
* Названия, состоящие из более чем одного элемента: «АТ», «Ca»,«’d»,«f%»
* Символ «-»
* Пробел, табуляция, перенос строки и прочие не видимые символы, которые программа игнорирует при загрузке короткого и длинного файлов

При загрузке допускается наличие имен в разных регистрах, так как потом они будут всё равно приведены к верхнему регистру:

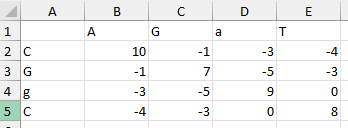


Если строки расположены в порядке, отличном от порядка столбцов, то программа распознает это и при загрузке файла переставит строки в том же порядке, что и столбцы.

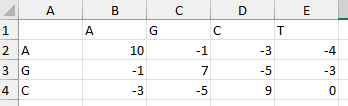
При загрузке пустые строки и столбцы матрицы будут проигнорированы:



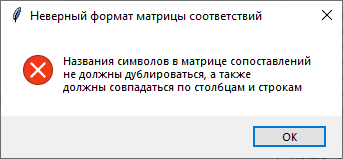
**НЕ ДОПУСКАЕТСЯ:**

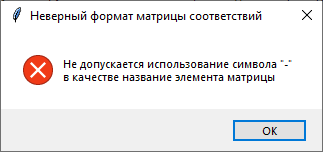
Наличие дублирования названий строк или столбцов вне зависимости от регистра 

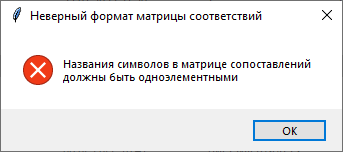
Использование не квадратных матрицы:



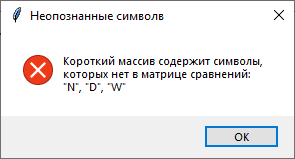
Типовые сообщения об ошибках:



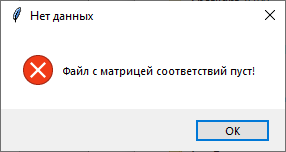




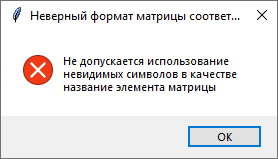
При возникновении любой из ошибок загрузка данного файла прерывается и результаты не сохраняются.



Возникает, если до этого был успешно загружен файл с короткой последовательностью и среди символов в коротком файле встретились те, что не присутствуют в качестве элементов матрицы сравнения. При появлении данной ошибки загрузка файла сравнений прерывается прерывается и данные не сохраняются.

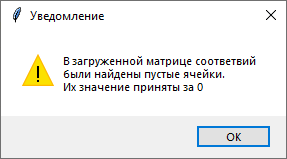


Возникает, когда в файле с матрицей соответствий нет ни одной строки и столбца данных. При появлении данной ошибки загрузка файла сравнений прерывается и данные не сохраняются.



Возникает, когда имя любого из элементов матрицы соответствий состоит из невидимого символа (пробел, табуляция и тд). При появлении данной ошибки загрузка файла сравнений прерывается и данные не сохраняются.

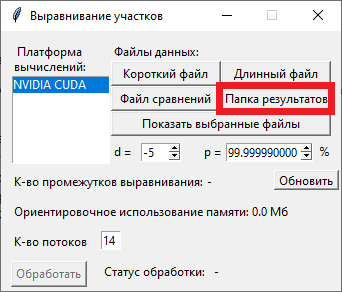
Типичное предупреждение



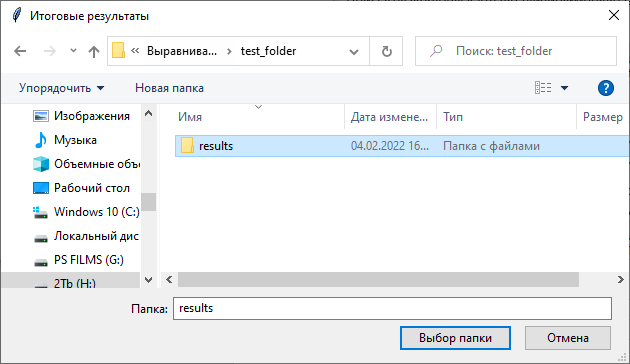
Возникает, когда в успешно загруженной матрице соответствий были найдены пустые ячейки. Это не вызывает ошибки и значения на месте пустых ячеек принимаются за 0.

Выбор папки сохранения результатов

Для это надо нажать на кнопку главной формы «Файл сравнений».



После чего указать в файловом диалоге путь до папки и нажать «Выбрать».



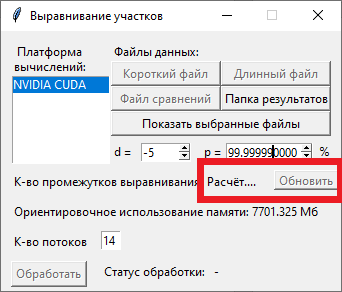
!!!

Если в процессе работы программы данная папка и любая из папок её содержащая будут удалены, то перед началом обработки все указанные в пути папки будут созданы автоматически.

Если в процессе обработки данных удалить папку результатов программа прекратит обработку без каких-либо сообщений об ошибке.

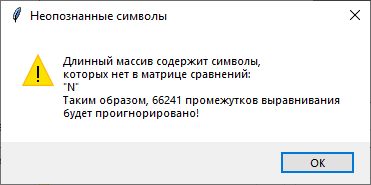
Проведение предобработки данных

Происходит после того, как были успешно загружены файл с короткой последовательностью, файл с длинной последовательностью и файл с матрицей сравнений. О процессе предобработки сигнализирует появление обновляющейся строки «Расчёт…» и неактивность 5-ти кнопок.



По завершение предобработки кнопки снова становятся активны.

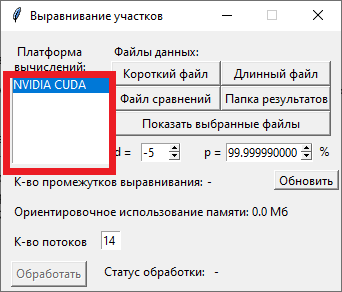
Типовое предупреждение



Возникает, когда после загрузки файла с короткой последовательностью, файла с длинной последовательностью и файла с матрицей сравнений обнаруживается, что среди символом в файле с длинной последовательностью имеются символы, что не присутствуют в качестве элементов матрицы сравнения. Не влияет на работу программы.

Необязательные действия

Задание платформы для параллельной обработки данных

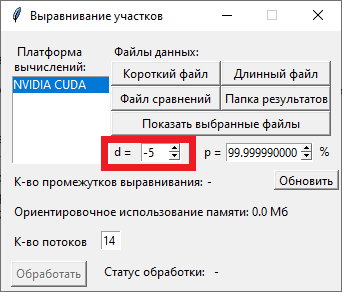


Отвечает за выбор набора аппаратных средств, на которых будет производиться вычисление. При наличие всего одного элемента в данном списке не стоит обращать на него внимаение.

В представленном случае предполагается использовать выдеокарту NVIDIA. Возможны другие варианты (Например «Intel(R) OpenCL HD Graphics» , а также два и более, если на компьютере например имеется процессор с встроенной графикой и дискретная видеокарта. В таком случае стоит выбирать наиболее производительную платформу.

Задать штрафной коэффициент **d**

Штрафной коэффициент вставки и удаления в алгоритмах выравнивания строк Хиршберга и Нидлмана-Вунша. Допустимые значения от -999 до 0.



Задать отсечку **p** биномиального распределения

Статистический параметр, задающий отсечку по количеству совпадений свертки совпадений по короткому и длинному файлу.

Отсечка по количеству совпадений расчитывается из параметра p следующим образом:

1.

Строится биномиальное распределение dist с параметрами:

n=arr\_short.length, p=alphabet.length и

где arr\_short – загруженный короткий массив

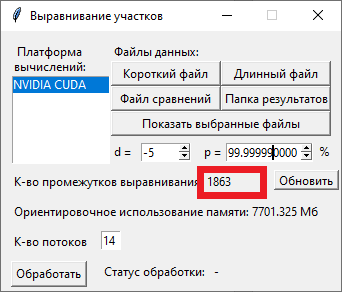
alphabet - массив названий колонок в матрице сравнений,

x – количества совпадений, для которых нужно посчитать вероятность при данных n и p

2.

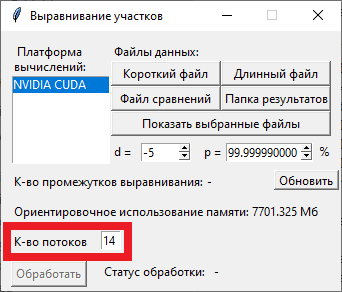
Находится индекс ***cutoff*** биномиального распределения, такой что

По завершение предобработки на месте записи «Расчёт…» указывается количества выравниваний, которые предстоит обработать программе при заданном значении **p**.



Количество промежутков выравнивания можно увеличить или уменьшить, соответственно увеличивая или уменьшая значение **p** с последующим нажатием на кнопку «Обновить».

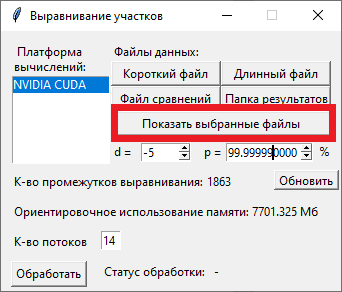
Задать количество потоков обработки



Обозначает количество поток, которые единовременно будут производить выравнивание строк. По умолчанию задаётся на два меньше, чем количество логических процессоров. Значение по умолчанию можно уменьшать, чтобы уменьшить пиковую нагрузку на процессор, но не стоит увеличивать, чтобы избежать перегрузки компьютера.

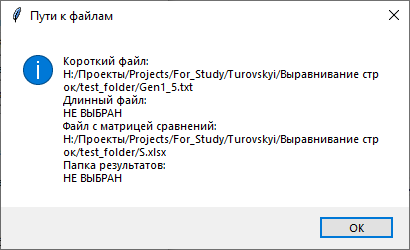
В зависимости от количества указанных потоков и размера короткого файла меняется значение «ориентировочное использование памяти», которое показывает, сколько при заданных параметрах конфигурации (будет показано дальше) программа максимум может использовать в сумме оперативной и видео памяти. Стоит заметить, что большую часть времени загрузка памяти намного меньше указанной величины, но про неё нельзя забывать, так как в случае недостатка памяти программа зависнет. Поэтому всегда стоит держать этот параметр меньше количества доступной в системе общей оперативной и видео памяти.

Просмотреть загруженные файлы



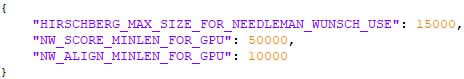
При нажатии показывает диалог, в котором приведены полные пути до файлов загруженных данных и путь к папке, в которую сохраняются результаты.

Пример диалога, когда выбрана только часть загружаемых файлов:



Файл конфигураций

Общий вид файла:



Данный файл должен находится в одной папке с запускаемым файлом и загружается при старте программы. При не верном значении параметров или имён полей берутся значения по умолчанию.

Значения полей:

HIRSCHBERG\_MAX\_SIZE\_FOR\_NEEDLEMAN\_WUNSCH\_USE – наибольшее количество элементов среди обоих последовательностей, поступающих на вход алгоритма Хиршберга, чтобы дальнейшее выравнивание происходило уже по методу Нидлмана-Вунша.

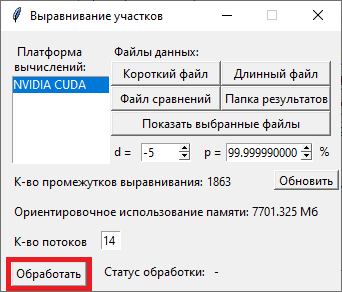
NW\_ALIGN\_MINLEN\_FOR\_GPU – наименьшая из длин двух последовательностей, полученных на входе алгоритма выравнивания Нидлмана-Вунша, при которой вычисления матрицы пути будет производиться на GPU. Влияние на память: O(n\*n)

NW\_SCORE\_MINLEN\_FOR\_GPU – наименьшее количество элементов среди обоих последовательностей, поступающих на вход суб-функции алгоритма Хиршберга по расчету прямого либо обратного счета (score), чтобы производить данные вычисления на GPU. Влияние на память как O(n).

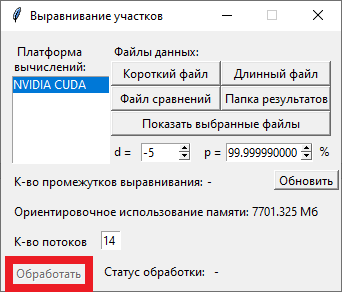
Обработка данных

Обработка данных становится доступна, когда загружены все файлы и выбрана папка для сохранения результатов.

Для начала обработки следует нажать на кнопку «Обработать».

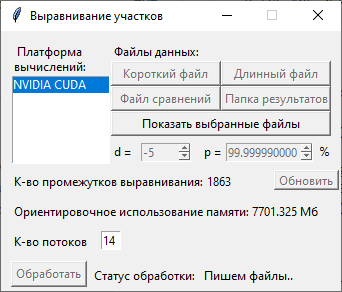


Если же файлы еще не загружены, то кнопка будет не активна.

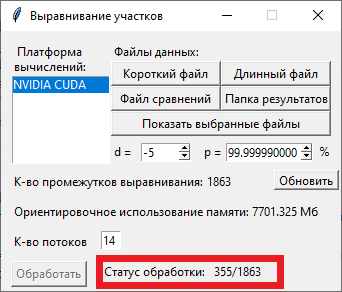


При нажатии на кнопку «обработать»

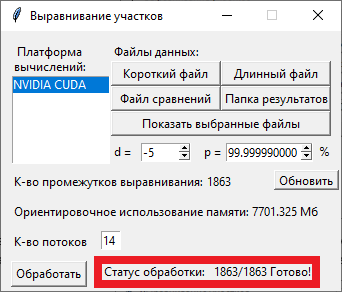
Сразу после нажатия на кнопку «Обработать» справа от статуса обработки появится надпись «Пишем файлы». Этот процесс займет некоторое время и важно, чтобы в этот момент ни один из файлов, в который программа пытается производить запись, не был открыт в системе



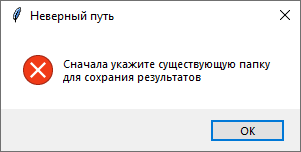
По завершении сохранения всех необходимых файлов справа от надписи «Статус обработки» начинает указываться в режиме реального времени количество обработанных промежутков выравнивания.



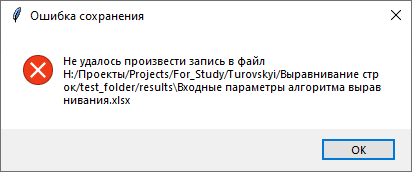
Завершение процесса выравнивания обозначается на форме надписью справа от количества обработанных промежутков выравнивания словом «Готово!»



Типовые ошибки:



Возникает, если при нажатии на активную кнопку «Обработать» ранее не была выбрана папка сохранения результатов. При возникновении данной ошибки Обработка прерывается.



Возникает, когда в начала обработки происходит попытка записи данных в какой-либо файл в папке с результатами. При возникновении данной ошибки Обработка прерывается.

Результаты обработки

Все результаты сохраняются в папку, которая была указана перед началом обработки.

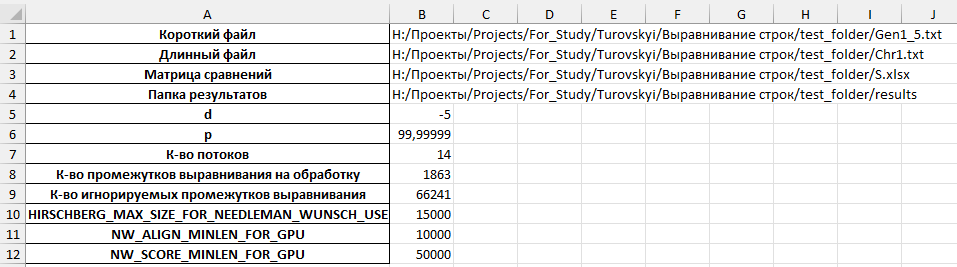
Всего есть 6 видов файлов:

1. Файл с входными параметрами алгоритма выравнивания
2. Файлы выравниваний – в количестве показанных перед обработкой промежутков выравнивания
3. Файл статистики
4. Файл биномиального распределения
5. Файл с указанием количества совпадений

Файл с входными параметрами алгоритма выравнивания

Обозначается как «Входные параметры алгоритма выравнивания.xlsx»

Указывает все параметры, указанные перед запуском алгоритма выравнивания, в том числе конфигурацию, загруженную из файла или использованную по умолчанию.



Файлы выравнивания

Представлены в количестве, соответствующем количеству показанного перед началом работы алгоритма количества промежутков выравнивания.

Обозначаются как «<ИНДЕКС\_НАЧАЛА\_ВЫРАВНИВАНИЯ\_ПО\_ДЛИННОЙ\_ПОСЛЕДОВАТЕЛЬНОСТИ>.txt»

На первой строке записано выравнивание короткой последовательности, на второй строке – выравнивание промежутка из длинной последовательности.

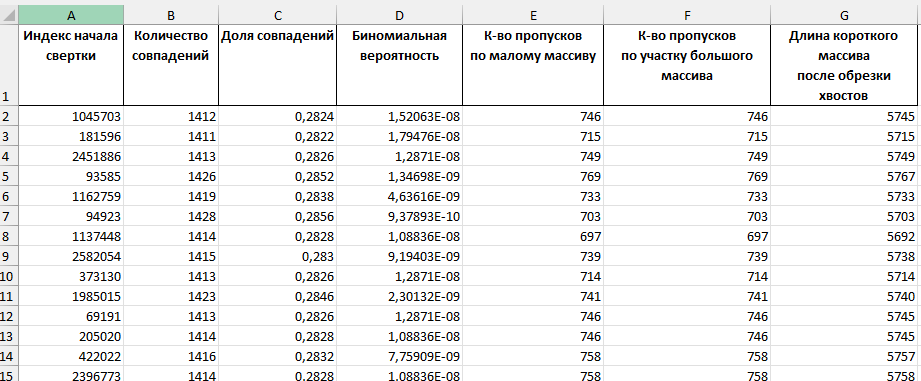


Файл статистики

Обозначается как «stats.xlsx»

Данный файл содержит в себе краткую информацию по каждому из промежутков выравнивания

1. Индекс начала свертки – Индекс в длинном массиве, начиная с которого считалось количество совпадений символов с коротким массивом.
2. Количество совпадений – общее количество совпавших символов на конкретном промежутке свертки
3. Доля совпадений – отношения количества совпавших символов на промежутке выравнивания к длине промежутка вырывания (что равно длине короткой последовательности)
4. Биномиальная вероятность –дискретное значение вероятности получения данного количества совпадений при n = <длина промежутка выравнивания> и p = 1/<Количество названий символов в матрице выравниваний>
5. Количество пропусков по малому массиву – общее количество пропусков, возникшее при использовании алгоритма выравнивания в коротком массиве и обозначенных в файле выравниваний как «-»
6. Количество пропусков по участку большого массива – общее количество пропусков, возникшее при использовании алгоритма выравнивания на промежутке из большого массива, по которому проводилась свёртка, обозначенных в файле выравниваний как «-»
7. Длина короткого массива после обрезки – длина короткого файла после выравнивания без учета символов «-» в начале и в конце выравненной последовательности



Файл биномиального распределения

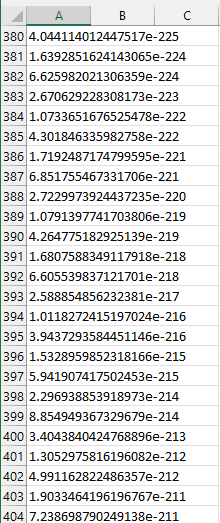
Обозначается как «Биномиальное распределение.xlsx»

Содержит в себе коэффициенты биномиального распределения для

n=<Длина промежутка выравнивания>, p=<К-во символов в матрице сопоставлений> и

Стоит заметит, что при значениях длинны короткой последовательности более нескольких сотен элементов в начале и конце файла с биномиальных распределением появляется элементы, равные 0. Связано это с ошибкой округления компьютера при расчете коэффициента.

Пример файла:



Файл с указанием количества совпадений

В данной файле указаны количества совпадений при свертке короткой и длиной последовательности в каждой возможной точке.

Специальное значение количества совпадений «-1» обозначает, что на данном промежутке выравнивания в длинной последовательности были встречены символы, которых не было среди символов в матрице сопоставлений, вследствие чего данный промежуток не участвовал в выравнивании.

