**Вариант 12**

**Цель работы:**

Целью данной лабораторной работы является обработка биологических данных, полученных из файлов формата GenBank. В рамках работы были выполнены следующие задачи:

1. Формирование единого файла GenBank из нескольких записей различных видов.

2. Вычисление и сортировка последовательностей по содержанию GC.

3. Извлечение и вывод белковых последовательностей, полученных в результате трансляции кодирующих областей.

**Задачи:**

1. Составить единый файл GenBank, содержащий не менее 10 различных кодирующих областей (CDS) из двух биологических видов.

2. Реализовать функцию для вычисления GC-составов последовательностей и вывести их в порядке возрастания.

3. Извлечь и вывести белковые последовательности, соответствующие кодирующим областям из файла GenBank.

**Инструменты и алгоритмы:**

Для выполнения лабораторной работы были использованы следующие инструменты и библиотеки:

- Библиотека: Biopython - для работы с биологическими данными в формате GenBank.

- Структуры данных: Список для хранения записей и кортежи для хранения информации о GC-содержании.

- Функции: Определены функции для выполнения каждой задачи:

1. calculate\_gc\_content — для вычисления GC-содержания последовательности.

2. extract\_protein\_sequences — для извлечения белковых последовательностей из CDS.

- Ввод/вывод: Использование стандартных функций print() для отображения результатов на экране.

Алгоритм работы программы

1. Формирование файла GenBank:

- С использованием Biopython считываются записи из двух файлов формата GenBank и объединяются в один файл combined\_species.gb.

2. Вычисление GC-составов:

- Читается файл combined\_species.gb, для каждой записи вычисляется GC-состав с помощью функции calculate\_gc\_content.

- Результаты сортируются по возрастанию GC-содержания и выводятся на экран.

3. Извлечение белковых последовательностей:

- Из файла combined\_species.gb извлекаются белковые последовательности из кодирующих областей (CDS) и выводится информация о каждой последовательности, включая идентификатор, описание, местоположение и саму трансляцию.

**Ошибки и их исправления:**

В процессе разработки программы были выявлены следующие ошибки:

1. Некорректная обработка пустых последовательностей: При вычислении GC-содержания не была предусмотрена проверка на пустую последовательность, что могло привести к делению на ноль.

- Исправление: Добавлена проверка на длину последовательности перед вычислением GC-содержания.

3. Отсутствие обработки ошибок при чтении файла: В случае отсутствия файла программа могла завершаться с ошибкой без информативного сообщения.

- Исправление: Добавлены блоки обработки исключений при открытии файлов, что позволяет избежать аварийного завершения программы.

**Выводы:**

В результате выполнения лабораторной работы была успешно реализована программа для обработки биологических данных в формате GenBank. Все поставленные задачи были выполнены: сформирован единый файл с кодирующими областями, рассчитаны и отсортированы GC-составы, а также извлечены белковые последовательности.