Отчет о летней практике
IG Container

Biocad, июль 2013

Дмитрий Кузьминов

Постановка задачи

- Требуется удобная структура для хранения данных:
  - нуклеотиды/аминокислоты
  - хранение аннотаций на каждый элемент
  - предлагаемая структура: бор
  - поддержка определенных операций (сл. слайд)

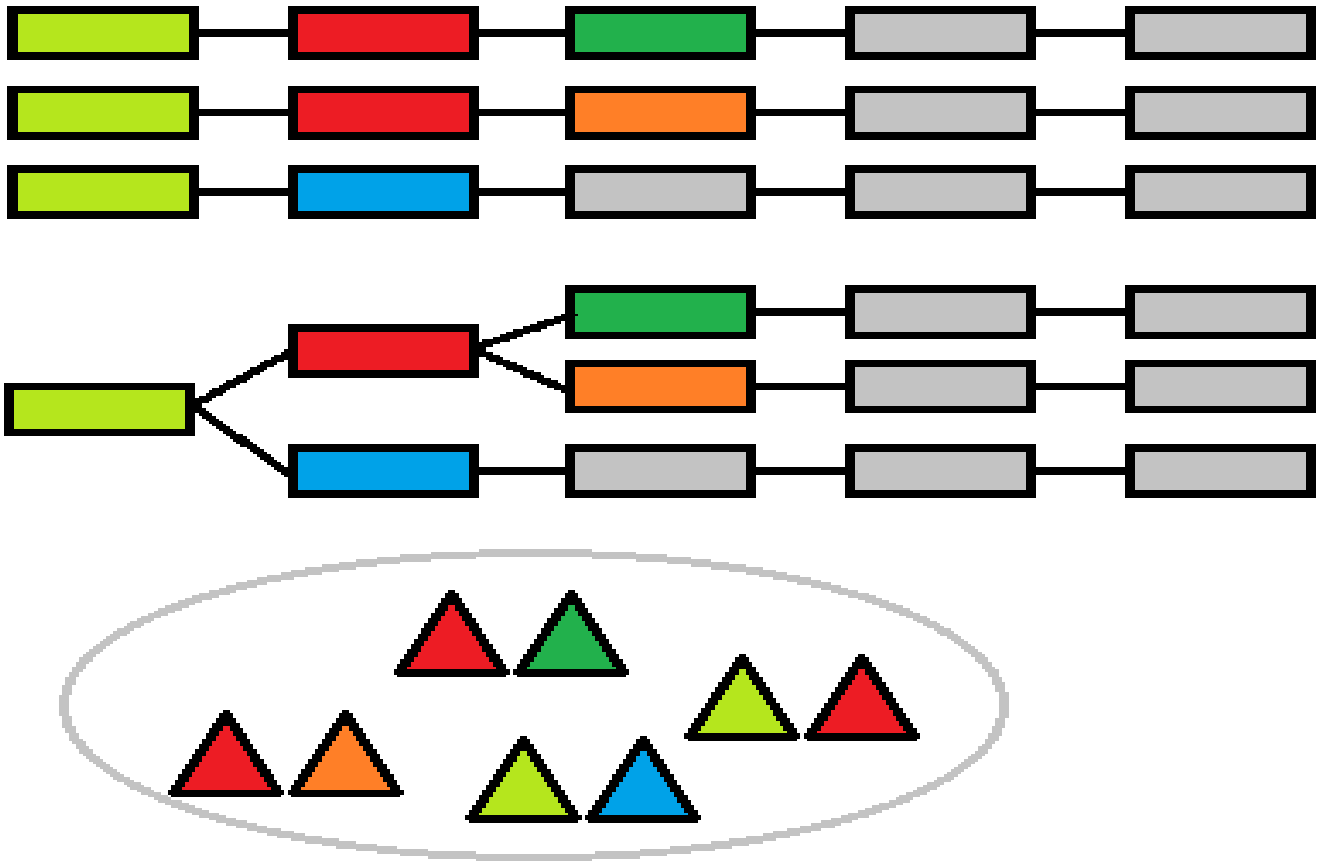
Постановка задачи

- Операции:
  - чтение из FASTA
  - сохранение в промежуточном формате (ну и чтение из него соответственно)
  - поиск паттерна
  - выравнивание паттерна на структуру

Предложенные идеи

- template<T>: нечто, не зависящее от хранимых данных
- Итераторы: инкапсуляция логики перебора элементов
- Тройка: бор–аннотации–k-меры

Реализация



Реализация (итоги на конец июля)

- Готовый модуль, поддерживающий операции:
  - чтение из FASTA
  - поиск паттерна
  - выравнивание (расстояние Левенштейна)

Реализация (итоги на конец июля)

- Операционные системы и технологии:
  - C++
  - разработка под Windows
  - портировано под Linux

Реализация (итоги на конец июля)

- Производительность:
  - быстрый поиск
  - выравнивание занимает порядка 15 секунд на реальных данных

Продолжение (итоги на конец августа)

- Рефакторинг кода
- Изменение концепций
  - возможность выбора vector<T> или deque<T>
  - дополнительные параметры шаблона
  - изменение структуры вершины в боре

Результаты (итоги на конец августа)

- Большая читаемость кода
- Ускорение операций в хранилище k-меров
- Большая гибкость при работе с памятью
- Странная потеря производительности на deque…

Планы

- Собрать в pipeline (командный интерпретатор)
- Выравнивание: различные матрицы замен
- Разобраться с deque
- Написание статьи