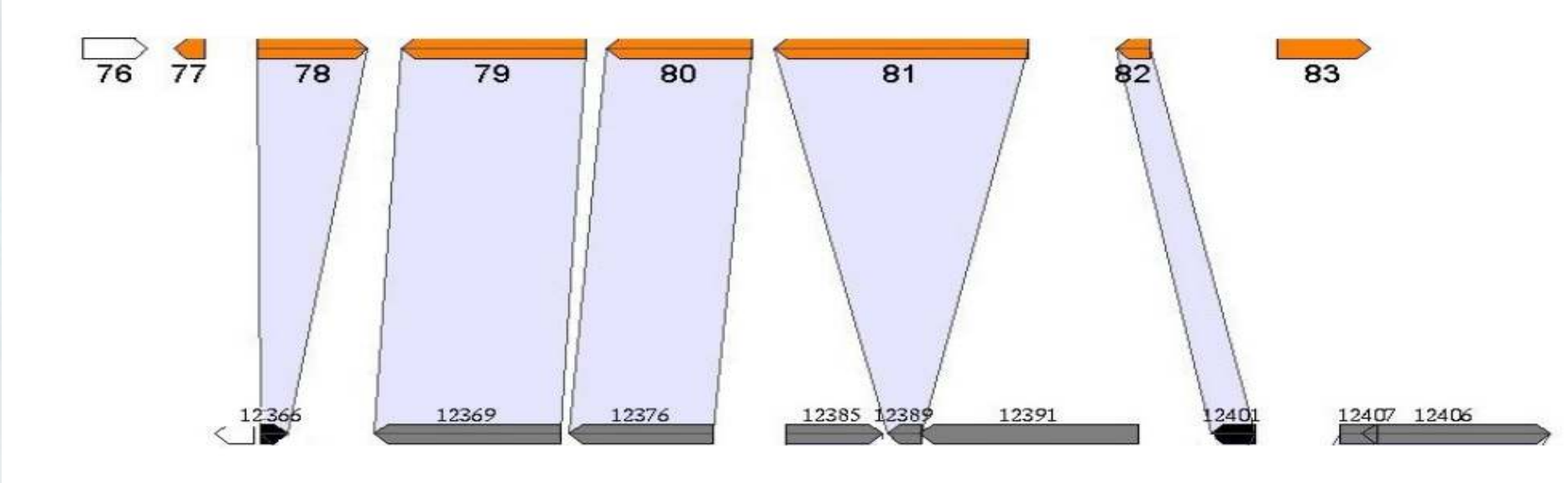
Multiple Global Alignment in C-Sibelia

Авдюхин Дмитрий
рук.: Минкин Илья

21 декабря 2013 г.

C-Sibelia

Полногеномное выравнивание



Multiple Alignment

- Вместо 2 последовательностей N
- Поиск гомологичных блоков легко масштабируется
- Проблема - глобальное выравнивание полученных блоков
- Существуют инструменты, решающие задачу, но они работают долго

Постановка задачи

- Провести анализ существующих глобальных выравнивателей
- Разработать более эффективный алгоритм для решения задачи множественного выравнивания
- Интеграция алгоритма в C-Sibelia

Изученные инструменты

- T-Coffie, Dialign-TX, MLagan, Clustal-Omega, MAFFT, MSARC
- Замерена их производительность
- Предназначены для работы в общем случае, а наши последовательности близки

Результаты

- Разработан алгоритм
  - Для каждого k-мера поиск похожих на него
  - Ищем близкие k-меры - min |p1 / len1 -- p2 / len2|
  - Битовое сжатие - сравнение k-меров за O(1)
  - Зафиксировали один нуклеотид для битовой маски - быстрый поиск близких k-меров
- Проведено сравнение алгоритма с существующими

Эксперименты

- 8 x ~200000 - справились MAFFT и MLagan
- Качество (mismatch cols / cols)
  - MLagan: 55810 / 258355
  - Наш: 56342 / 260357
- Время работы
  - MLagan: 88 секунд
  - MAFFT: 245 секунд
  - Наш: 1.87 секунды

TODO

- Нормальное прогрессивное выравнивание
- Более устойчивый
- Обработка N - произвольного нуклеотида
- Интеграция в C-Sibelia
- Сравнение C-Sibelia с существующими полногеномными выравнивателями