Проекты для бакалавров

1. Архивация FASTQ файлов (Андрей)

Секвенирование — процесс “чтения” небольших фрагментов ДНК. Размеры этих фрагментов, как правило, составляют от 100 до 10000 нуклеотидов, в то время как длина всей цепи ДНК может составлять миллионы и миллиарды нуклеотидов в зависимости от организма. В силу несовершенства современных технологий читать ДНК целиком никто не умеет. В результате секвенирования получаются огромные файлы, в которых записаны миллионы прочитанных фрагментов. Одним из самых популярных форматов является простой текстовый формат FASTQ. В этом проекте предлагается попробовать разработать и реализовать эффективные алгоритмы, которые учитывали бы особенности данного формата и природу данных секвенирования.

2. Архивация данных, алгоритмы класса LZ (Андрей)

Алгоритмами архивации можно заниматься и без привязки к биоинформатике. В качестве проекта предлагается реализовать некоторые алгоритмы класса LZ, а также попробовать адаптировать его под какой-нибудь конкретный тип данных (например, исходного кода).

3. Первичный анализ данных сиквенирования (Андрей, Лёша)

Одной из актуальных задач является первичный анализ данных секвенирования. Пользователь программы должен получить достаточное количество статистик и графиков, которые помогли бы ему оценить качество проведенного эксперимента и предсказать потенциальные проблемы использования этих данных.

4. Визуализация сборок (потенциально можно 2 человека, Андрей, Лёша)

Современные технологии секвенирования позволяют считывать достаточно короткие фрагменты ДНК, которые не поддаются осмысленному анализу. Для “склейки” коротких фрагментов в более длинные последовательности используют специальные программы, называемые геномными ассемблерами, или сборщиками. В результате сборки мы получаем более длинные последовательности, которые поддаются анализу и могут быть использованы в дальнейших биологических исследованиях. Однако, сборки, как правило, содержат достаточно большое количество последовательностей разной длины, работать с которыми не всегда удобно. Для облегчения их анализа, предлагается реализовать интерактивные визуализатор геномных сборок.

5. Кластеризация Хэмминг графа, построенного по Ig-Seq ридам

Описание: аккуратное построение репертуара антител является важным предварительным этапом различных задач иммуноинформатики. Современные платформы секвенирования (например, Illumina MiSeq) позволяют получить риды, целиком покрывающие

вариабельный участок антитела (самую важную его часть, отвечающую за связывание с антигенами, т.е. потенциально вредными белками), но допускают ошибки (например, замена одного нуклеотидного остатка на другой). Таким образом, задача построения репертуара с помощью секвенирования сводится к отделению ошибок секвенирования от натуральных вариаций. Для этих целей мы используем Хэмминг граф, построенный по целым ридам. В рамках данного проекта предлагается разработать метод кластеризации Хэмминг графа, позволяющий выделить ошибки в ридах и отделить их от натуральный вариаций на уровне структуры графа.
Контакты: safonova.yana@gmail.com (Яна Сафонова)

6. Де Брюйн граф на идеальных хешах: упрощение и коррекция ридов (Антон Б)

Де Брюйн граф это одна из базовых структур данных, используемых для анализа и сборки геномов. Эффективные его хранение и обработка являются важной для биоинформатики задачей. Один из подходов к этой задаче это идеальное хэширование: легковесное отображение объектов (вершин графа) в их номера (числа от 1 до n). Задача проекта: научиться использовать эту структуру данных для реализации стандартных операций с графом де Брюйна: упрощение графа, коррекция ридов. Нужно будет работать с готовой реализацией этой структуры, дорабатывать её и использовать. Язык C++.

7. Striped Smith-Waterman. Векторизация на AVX / AVX2 (Антон К.)

Существуют эффективные реализации алгоритма Смита-Ватермана, использующие инструкции набора SSE2 (см. например, http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0082138) , однако нам не известны аналогичные реализации с использованием AVX / AVX2. В рамках данного проекта предлагается сделать подобную реализацию.

8. Эффективная параллельная сортировка к-меров: parallel merge sort, radix sort и т.п. (Антон К.)

9. Прототипирование аминокислотных последовательностей антител (Кира)

Антитела - это белки, вырабатываемые иммунной системой в ответ на появление в организме чужеродных веществ, называемых антигенами. Каждое антитело образовано двумя парами идентичных аминокислотных цепей – *легкой* и *тяжелой*. Каждая цепь состоит из *константного* и *вариабельного* регионов.

Аминокислотная последовательность константного региона является приблизительно одной и той же для каждого из типов тяжелой или легкой цепи (IgA, IgE, IgD, IgG, IgM –

для тяжелой цепи, κ и λ – для легкой цепи). Вариабельный же регион определяет антигенную специфичность антитела; он включает в себя четыре *каркасных* (*framework*) и три *гипервариабельных* участка (*complementarity determining regions*, *CDRs*).

При восстановлении аминокислотной последовательности антитела по набору масс-спектров гипервариабельные ее участки, представляющие наибольший интерес, могут быть получены лишь при помощи методов *de novo* севенирования. В то же время, для константного региона и каждого из каркасных участков его «прототип» может быть сформирован на основе результатов идентификации масс-спектров посредством поиска в базе данных. К их построению и сводится задача *прототипирования* аминокислотной последовательности антитела.

Соответствующая алгоритмическая задача будет сформулирована в терминах строк в алфавите из 20 стандартных аминокислот.