Современная биоинформатика: эффективные алгоритмы обработки данных омиксных технологий

Вяткина Кира Вадимовна

СПбАУ РАН им. Ж.И. Алферова
Институт трансляционной биомедицины СПбГУ
СПбГЭТУ «ЛЭТИ»
ННЦН – филиал ФГБУ «НМИЦ психиатрии и наркологии им. В.П. Сербского»

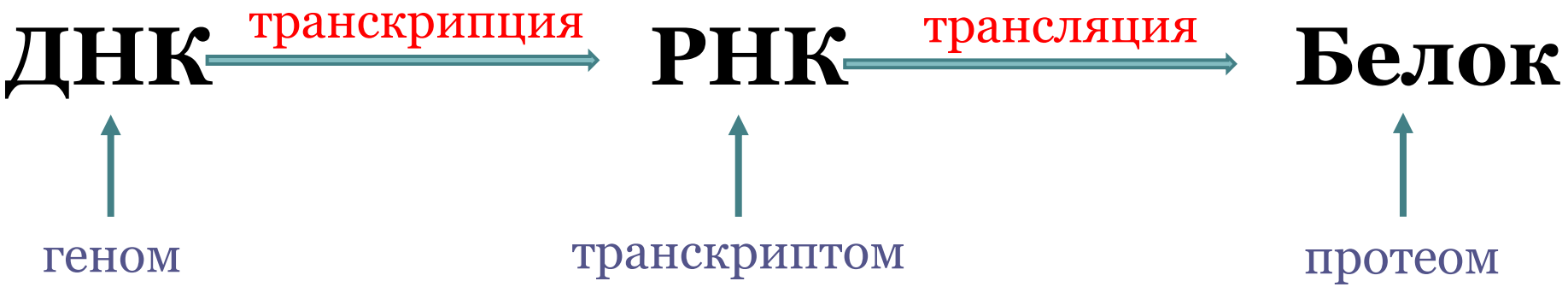
vyatkina@spbau.ru, kira.vyatkina@gmail.com

Семинар «Информатика и автоматизация», СПИИРАН, 25 ноября 2022 г.

Биоинформатика

- Алгоритмические и вычислительные задачи возникают в самых разнообразных областях биологии
- Их решением занимаются специалисты в области биоинформатики
- В частности, биоинформатические подходы широко применяются при решении задач «омиксных» направлений
  - Геномика
  - Транскриптомика
  - Протеомика
  - Метаболомика

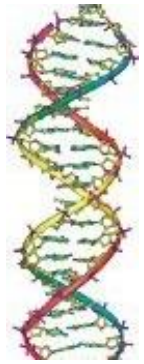
Центральная догма молекулярной биологии



ДНК: дезоксирибонуклеиновая кислота
РНК: рибонуклеиновая кислота

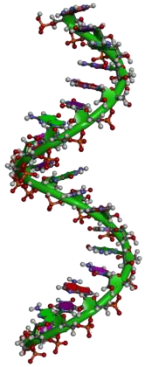
ОМИКсные технологии

Вычислительная геномика



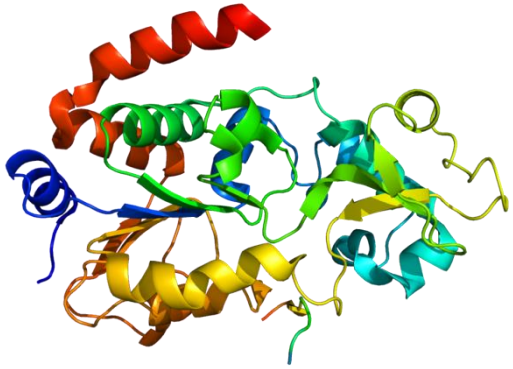
ДНК

Вычислительная транскриптомика



РНК

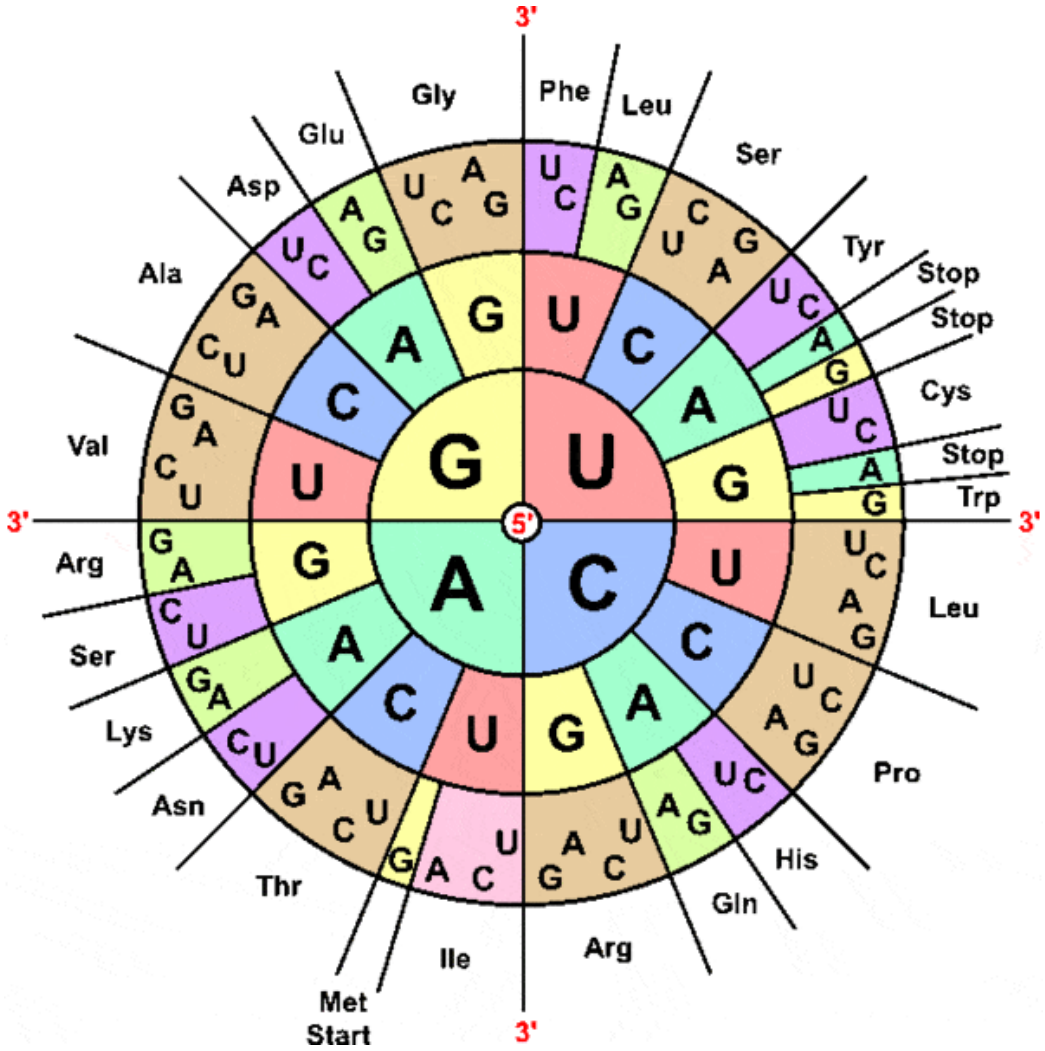
Вычислительная протеомика



Белок

Генетический код

- ДНК: 4 нуклеотида
  - A – аденин
  - C – цитозин
  - G – гуанин
  - T – тимин
- РНК: 4 нуклеотида
  - A – аденин
  - C – цитозин
  - G – гуанин
  - U – урацил
- Белок: 20 аминокислот



Анализ последовательностей

- Установление первичной структуры
  - Секвенирование
- Выравнивание последовательностей
- Задачи геномики
  - Сравнительная геномика
  - Построение филогенетических деревьев
  - Поиск мотивов
  - CG-состав

Секвенирование

- Определение нуклеотидной последовательности ДНК или РНК
- Определение аминокислотной последовательности белка
- Иными словами: установление первичной структуры ДНК, РНК или белка

Геном

- С вычислительной точки зрения: строка в алфавите из четырех букв (нуклеотидов)

- Длина генома
  - человека: ~3 миллиарда пар нуклеотидных оснований (base pairs, bp)
  - кишечной палочки: ~4,6 миллионов пар нуклеотидных оснований
  - амебы: ~670 миллиардов пар нуклеотидных оснований

Сборка генома

- Современные технологии не позволяют прочитать геном целиком
- Вместо этого может быть получен набор относительно коротких фрагментов генома – «прочтений» (reads)
  - Illumina MiSeq: 50 bp, 150 bp, 250 bp, 300 bp
- Задача сборки генома (genome assembly) заключается в восстановлении его последовательности по набору прочтений

Задача сборки генома

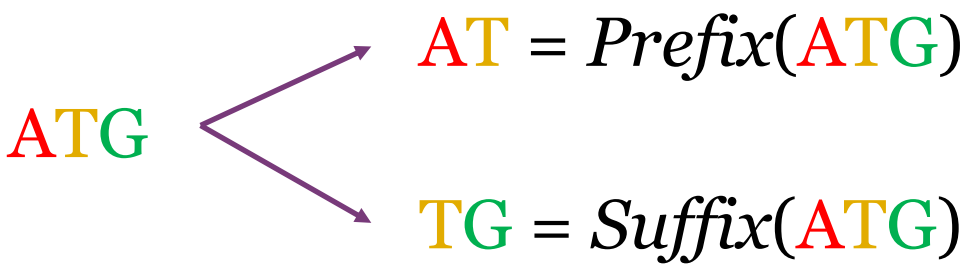
- Дан набор фрагментов генома длины k (k-меров)
- Требуется восстановить геномную последовательность
- **TAATGGGATGCCATGTT**
  - **AAT, ATG, ATG, ATG, CAT, CCA, GAT, GCC, GGA, GGG, GTT, TAA, TGC, TGG, TGT**

  - **TAATGGGATGCCATGTT**

Обозначения

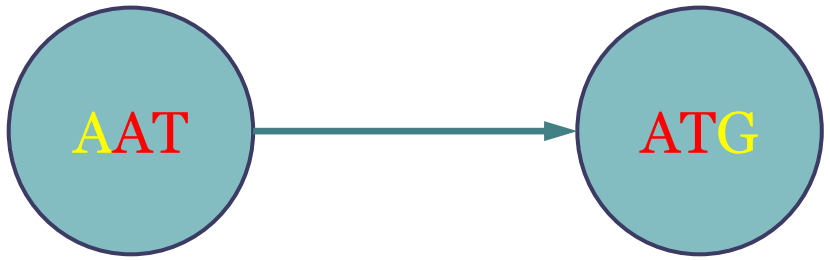
- Для каждого k-мера его *префикс* образован первыми k-1 нуклеотидами, а *суффикс* – последними k-1 нуклеотидами

- k=3



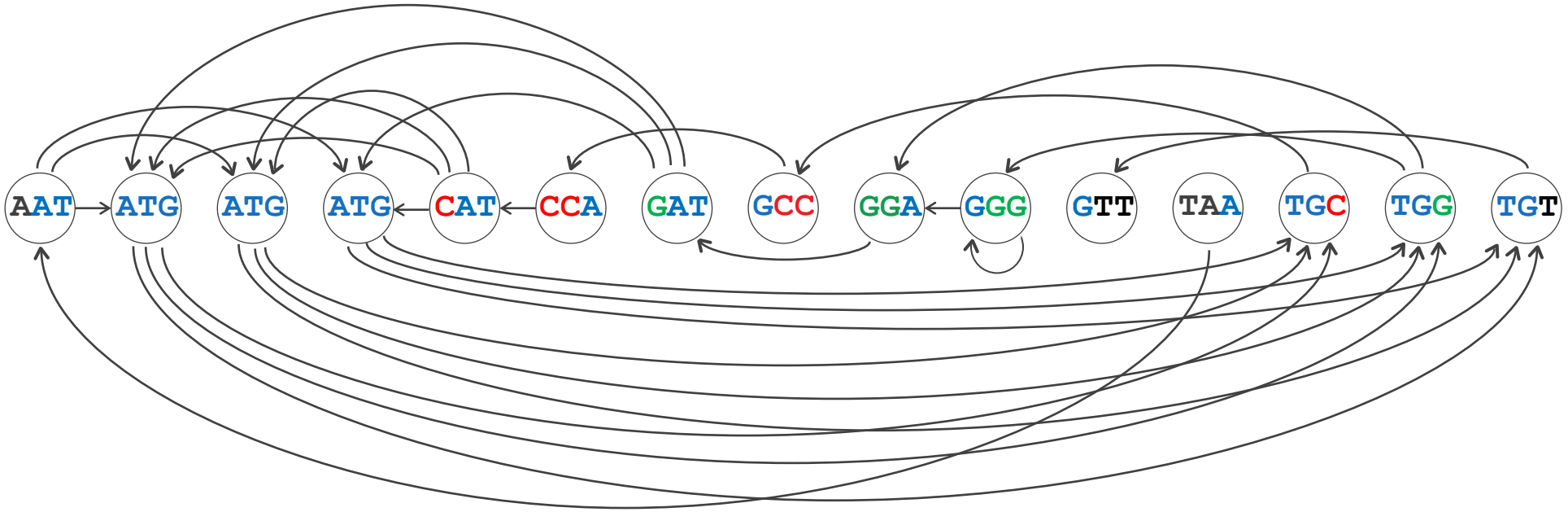
Граф перекрытий (overlap graph)

- Сопоставим каждому k-меру вершину графа перекрытий
- Соединим ориентированными ребрами вершины, помеченные k-мерами a и b, если $Suffix(a)=Prefix(b)$



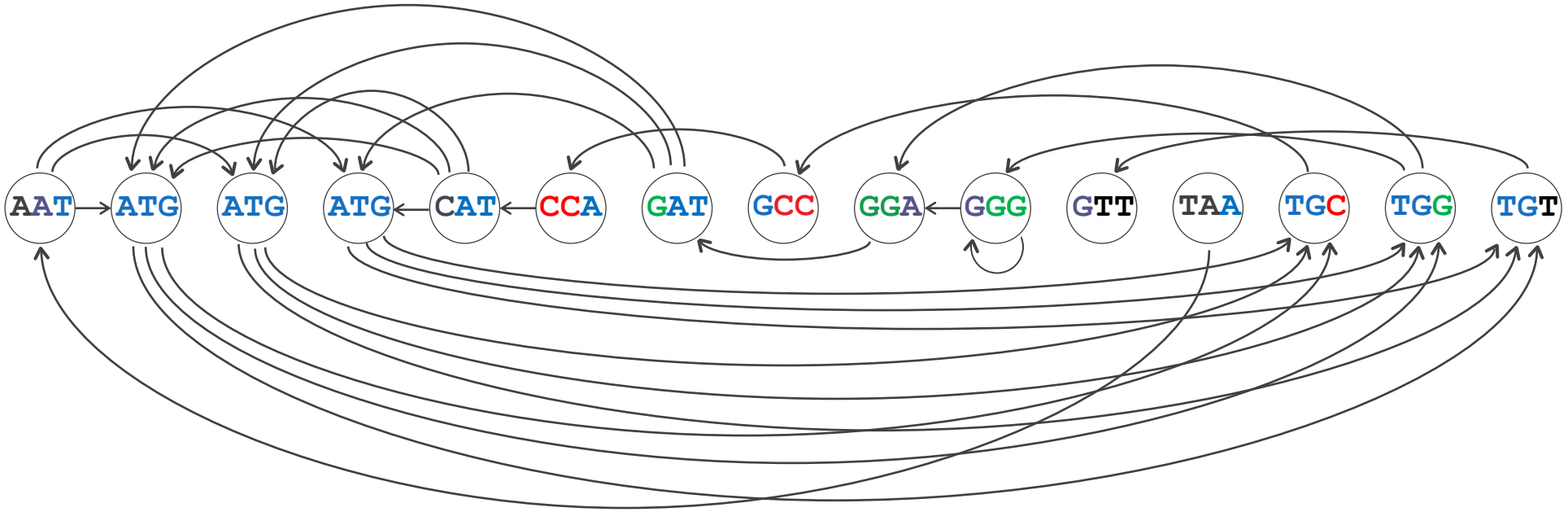
Граф перекрытий (overlap graph)

- Сопоставим каждому k-меру вершину графа перекрытий
- Соединим ориентированными ребрами вершины, помеченные k-мерами a и b, если $Suffix(a)=Prefix(b)$



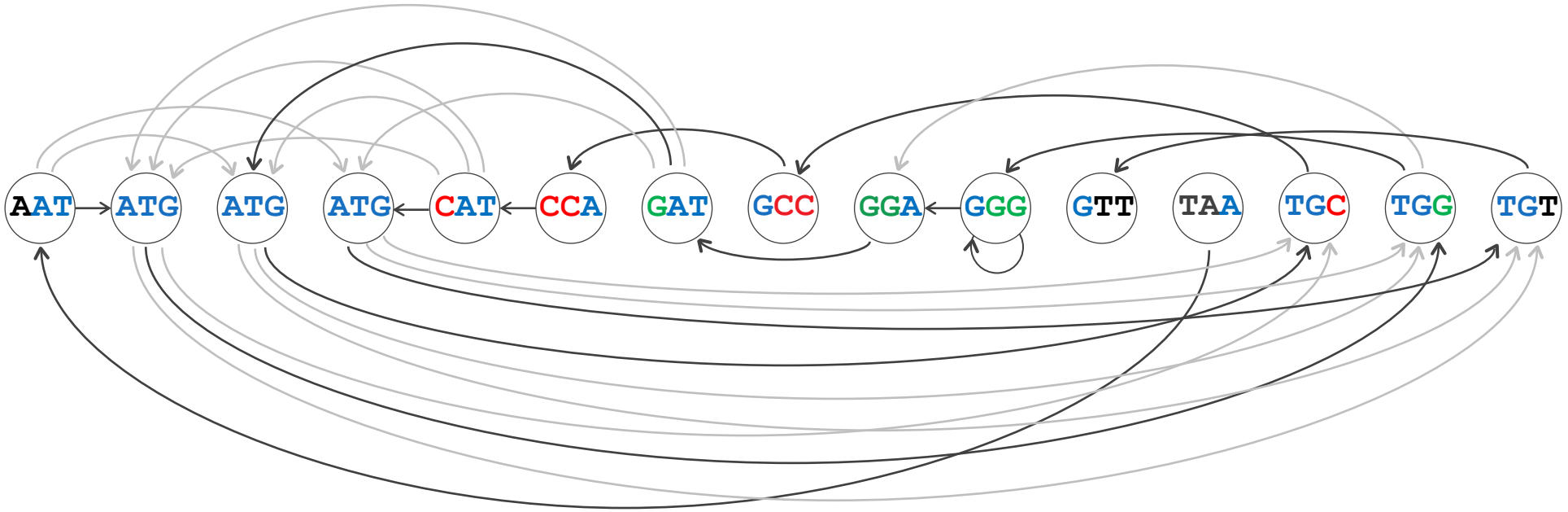
Граф перекрытий (overlap graph)

- Гамильтонов путь проходит через каждую вершину графа ровно один раз
- Он определяет геномную последовательность



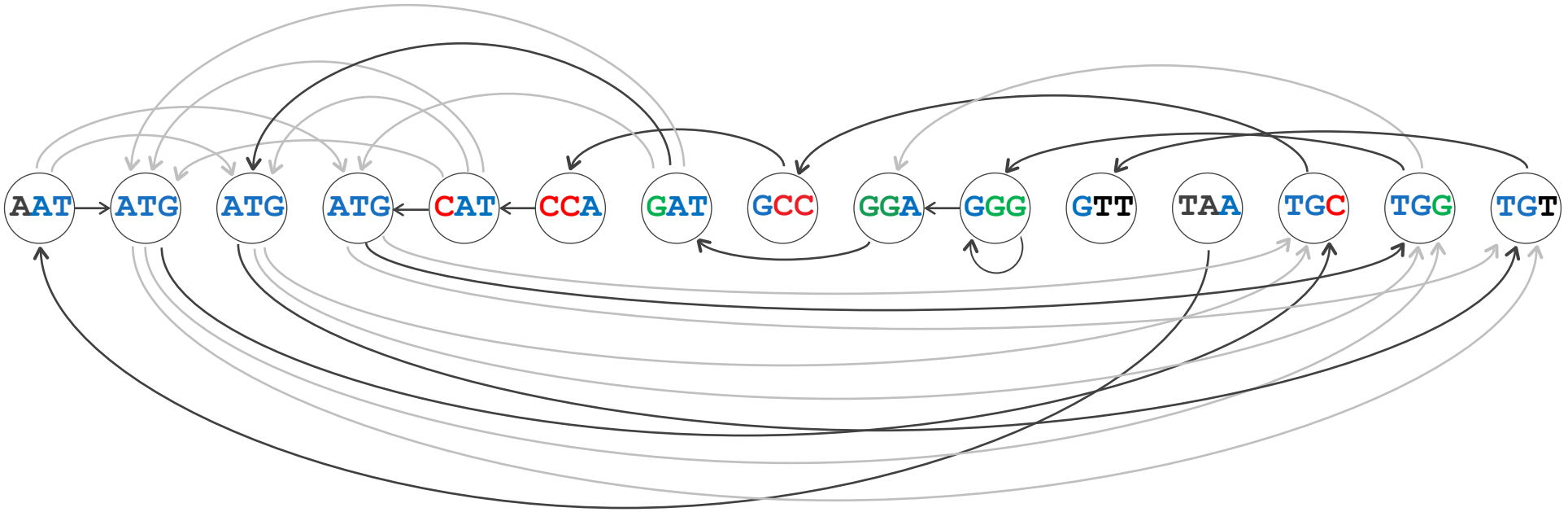
Граф перекрытий (overlap graph)

- **TAATGGGATGCCATGTT**



Граф перекрытий (overlap graph)

- **TAATGCCATGGGATGTT**

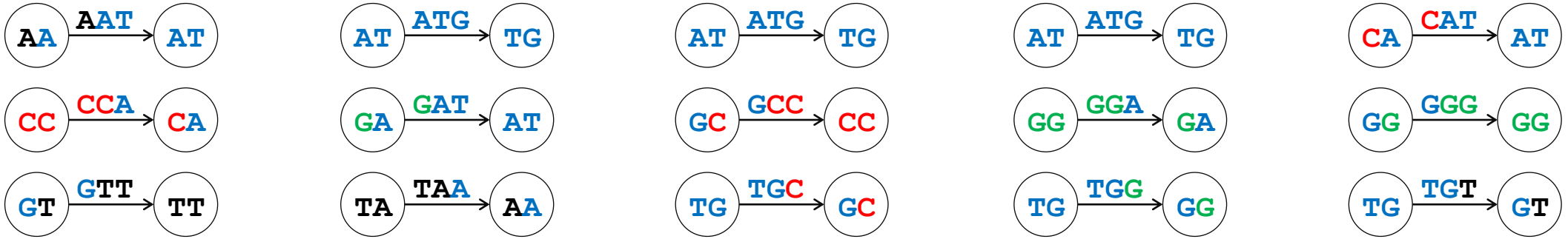


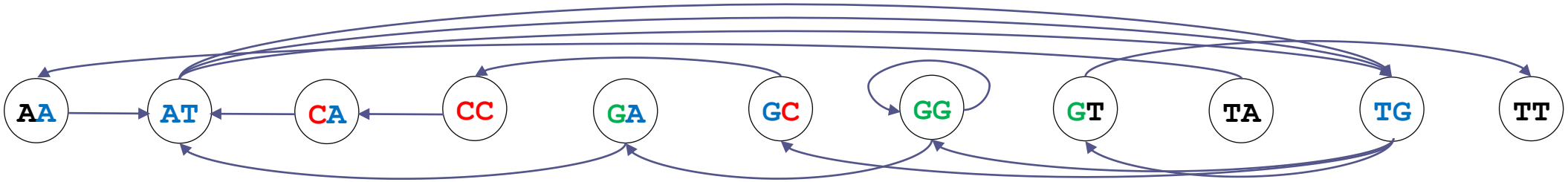
Граф де Брюйна (De Bruijn graph)

- Для каждого k-мера сгенерируем помеченное им ориентированное ребро, а начальную и конечную вершину этого ребра пометим его префиксом и суффиксом
- Склеим вершины, помеченные одинаковыми (k-1)-мерами

Граф де Брюйна (De Bruijn graph)

- **TAATGGGATGCCATGTT**
  - **AAT , ATG , ATG , ATG , CAT , CCA , GAT , GCC , GGA , GGG , GTT , TAA , TGC , TGG , TGT**





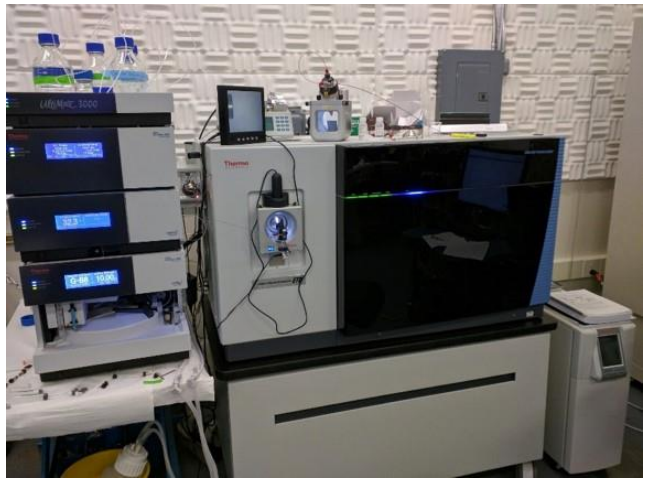
Сложность решений

- Если P≠NP, задача построения гамильтонова пути не может быть решена за полиномиальное время

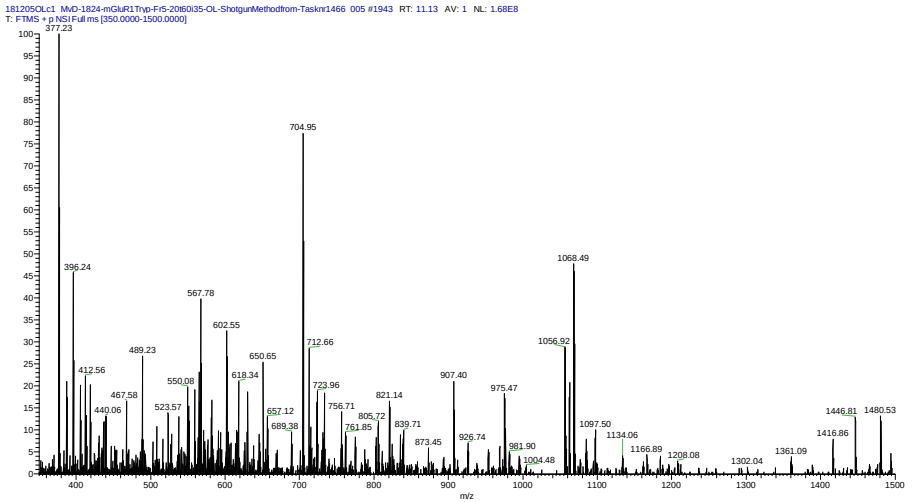
- Эйлеров путь может быть построен за линейное время

Масс-спектрометрия

- Масс-спектрометр измеряет отношение m/z массы к заряду ионизированных молекул
- Результат измерений: масс-спектр



https://proteomics.ucdavis.edu/thermo-fusion-lumos-2016



Методы анализа

- Панорамный анализ (shotgun proteomics)
  - Цель: идентифицировать как можно больше белков, содержащихся в образце

- Направленный анализ (targeted analysis)
  - Содержатся ли в образце определенные белки?
  - В каких количествах они присутствуют?

Анализ белков и пептидов

NQEL
NQEL
NQEL
NQEL
NQEL
NQEL
NQEL
NQEL
NQEL

Фрагмент	Масса
N	114
NQ	242
NQE	371
QEL	370
EL	242
L	113

Идентификация белков и пептидов

Фрагмент	Масса
N	114
NQ	242
NQE	371
QEL	370
EL	242
L	113

Методы идентификации белков и пептидов

- Поиск в базе данных
  - Database search
- Поиск в спектральных библиотеках
  - Spectral library search
- *De novo* секвенирование
  - *De novo* sequencing
- Поиск в базе данных с использованием тегов пептидных последовательностей
  - Sequence tag search

Методы идентификации белков и пептидов

- Поиск в базе данных
  - Имеется база данных аминокислотных последовательностей
  - Цель: найти последовательность, которая лучше всего объясняет экспериментальный масс-спектр
- *De novo* секвенирование
  - Установление последовательности только по масс-спектру
  - Как правило, используются *спектральные графы*

Поиск в базе данных

- База данных белковых или пептидных последовательностей генерируется на основе генома
- Экспериментальный масс-спектр сопоставляется с теоретическими, сгенерированными для белков или пептидов из базы данных
- Наилучшее соответствие выбирается с использованием заданной подходящим образом оценочной функции

De novo секвенирование

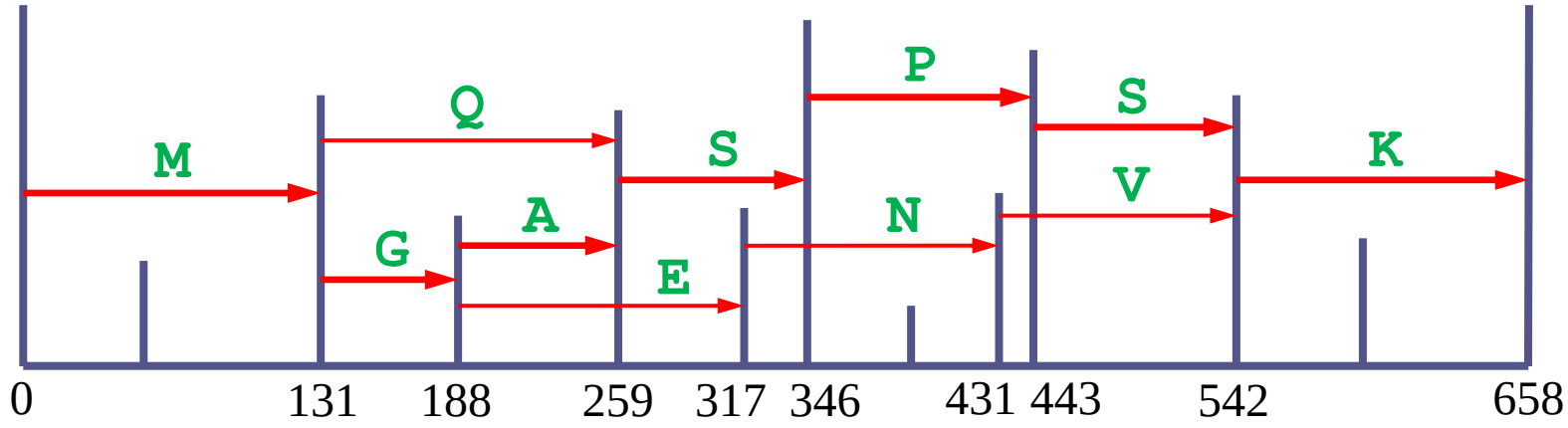
- Не предполагает никаких априорных знаний об анализируемом белке

- Существующие алгоритмы используют
  - Спектральные графы
    - Основная идея: выявление последовательностей однотипных ионов (ion ladders)
  - Метод динамического программирования

De novo секвенирование

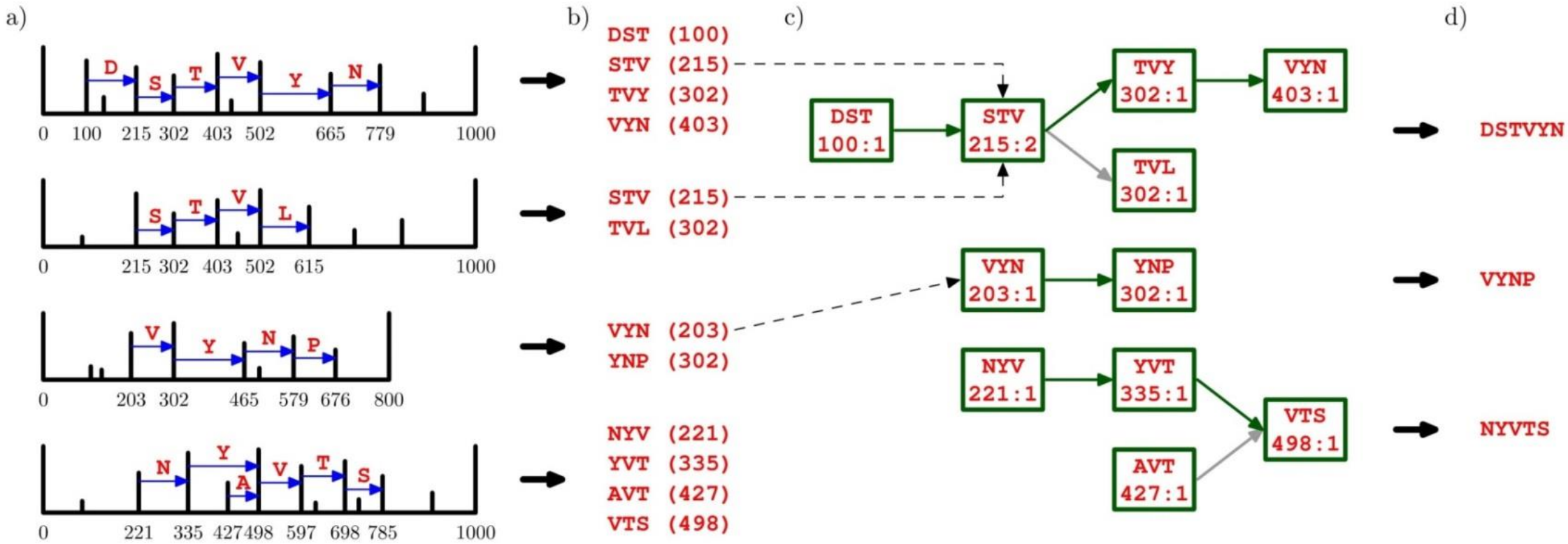
- Bottom-up:
  - Lutefisk (Taylor et al., 1997)
  - Sherenga (Dancik et al., 1999)
  - PEAKS (Ma et al., 2003)
  - PepNovo (Frank et al, 2005)
  - pNovo (Chi et al., 2010)
  - Novor (Ma, 2015)
  - DeepNovo (Tran et al., 2017)
  - DeepNovo-DIA (Tran et al., 2019)
- Top-down:
  - Twister (Vyatkina et al., 2015, 2016, 2017)
    - Was adopted to the bottom-up case

Спектральный граф



MGASPSK

Алгоритм Twister



Актуальные направления исследований

- Метагеномика/метапротеомика
- Геномика/протеомика одной клетки
- Хемопротеомика/разработка лекарственных средств
- Обработка омиксных данных, получаемых с применением новых технологий