Создание классификатора для определения эффекторных белков системы секреции VI грамотрицательных бактерий

Иван Петрушин
Иркутский государственный университет
Лимнологический институт СО РАН

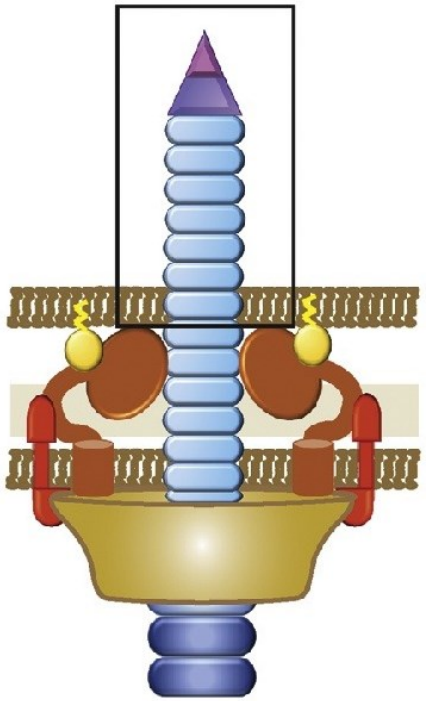
Предсказание свойств белков *in silico*

Популярный способ отбора

- Аннотация (предсказание функций) генов
  - GeneMark, Glimmer
- Поиск сайтов в белках
  - SignalP
  - TargetP
- Анализ гидрофобности

Type VI secretion system

- Распространена у патогенных бактерий
- Компоненты-белки консервативны
- Секретирует широкий спектр белков
  - Токсины
  - Гидролазы
  - Металлофоры (получение ионов металлов)



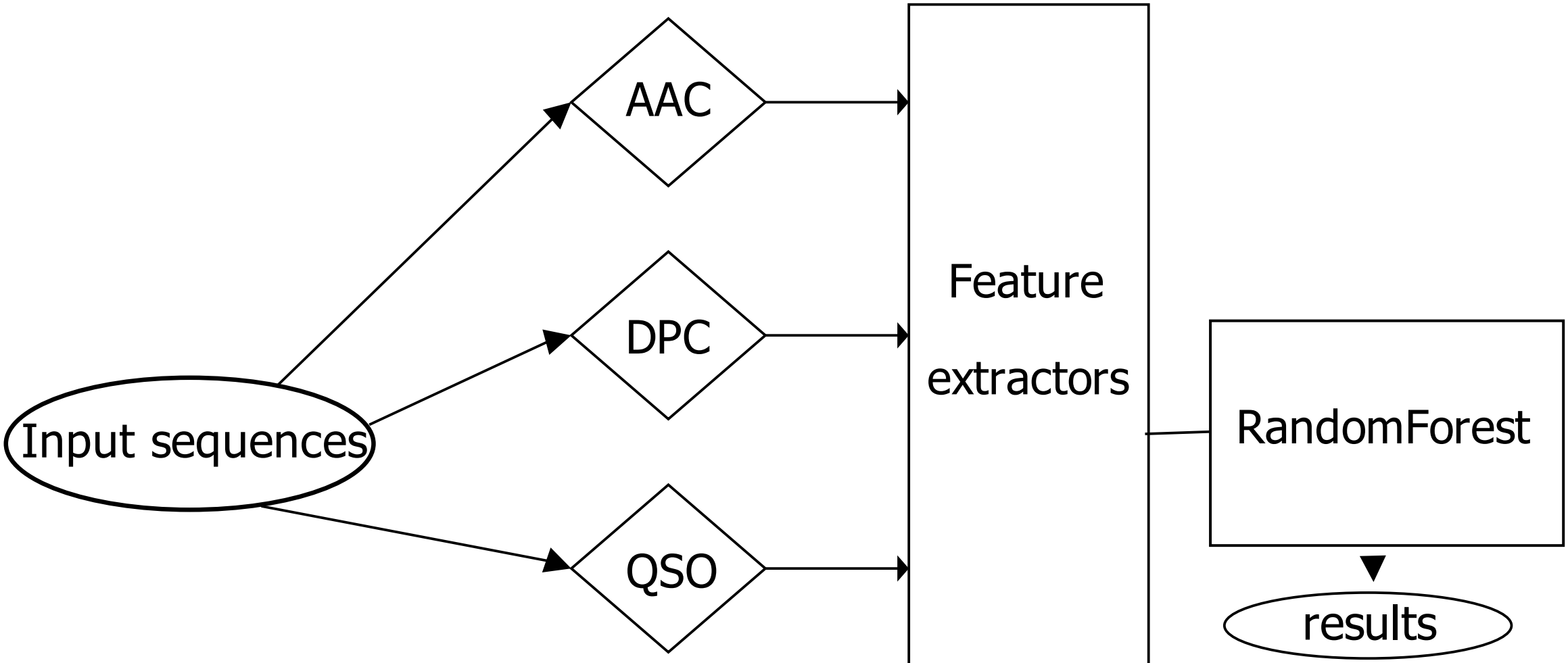
Предсказание класса (эффектор)

Исходные данные – последовательность аминокислот

Признаки:

- Последовательность
  - AAC (частотность аминокислот)
  - DPC (частота биграммов)
  - QSO (quasi-sequence-order)
- Эволюционные (PSSM, BLOSUM)
- Физико-химические
- Сравнение с базами белков

Архитектура модели



Архитектура модели

Вычисление ряда признаков может быть сложным

Position **S**pecific **S**ubstitution **M**atrix – 5 минут (!)

Используется только первая группа признаков

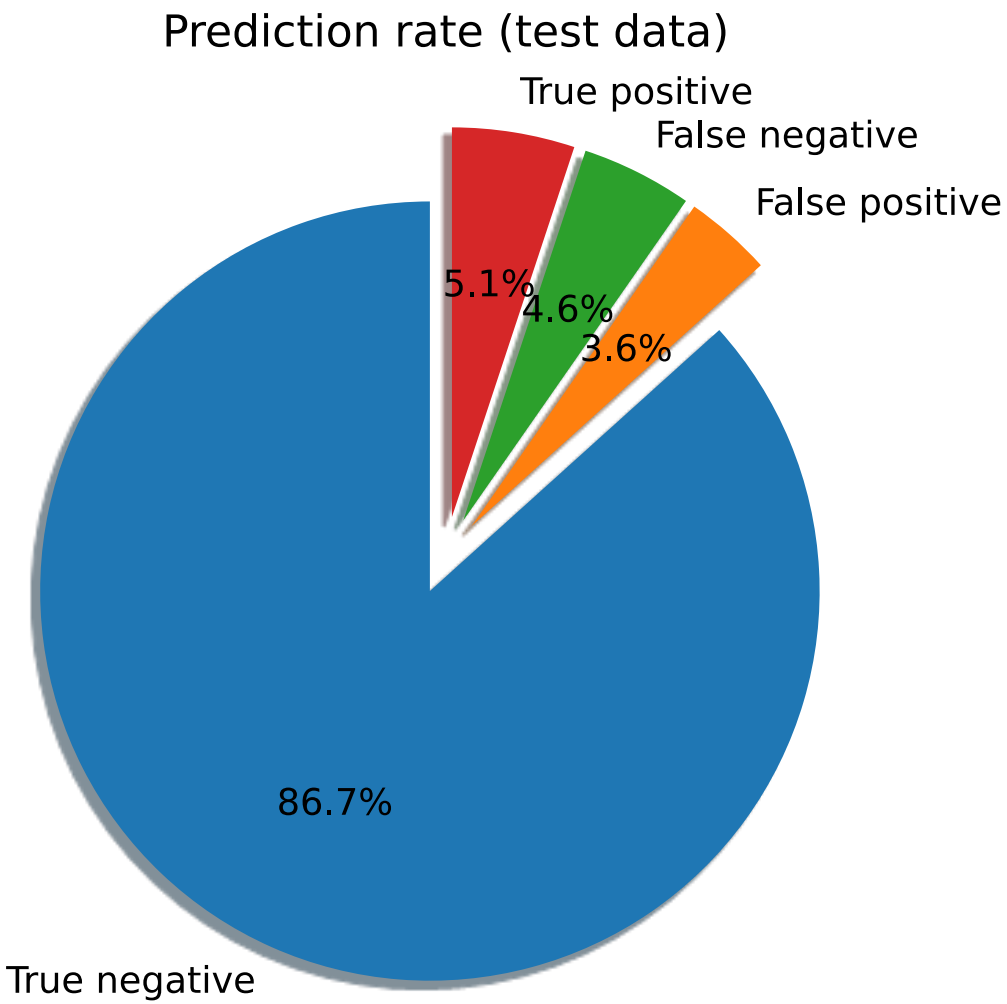
Два уровня: SVC для AAC, DPC, QSO – базовые модели

Random Forest – финальная модель

Модульная архитектура (ансамбли) позволяет добавлять признаки и тестировать независимо

Тестирование модели

- Проблема – данные не сбалансированы
- Ансамблевый подход затрудняет кросс-валидацию
- Значения TN/TP/FP/FN зависят от разбиения



Планы

- Расширить обучающую выборку
- Реализовать остальные признаки
- Использовать многопоточность