Институт биоинформатики

Вариация в свете функции

Выполнили:
Иван Ревегук
Алевтина Корешова
Диана Кондинская

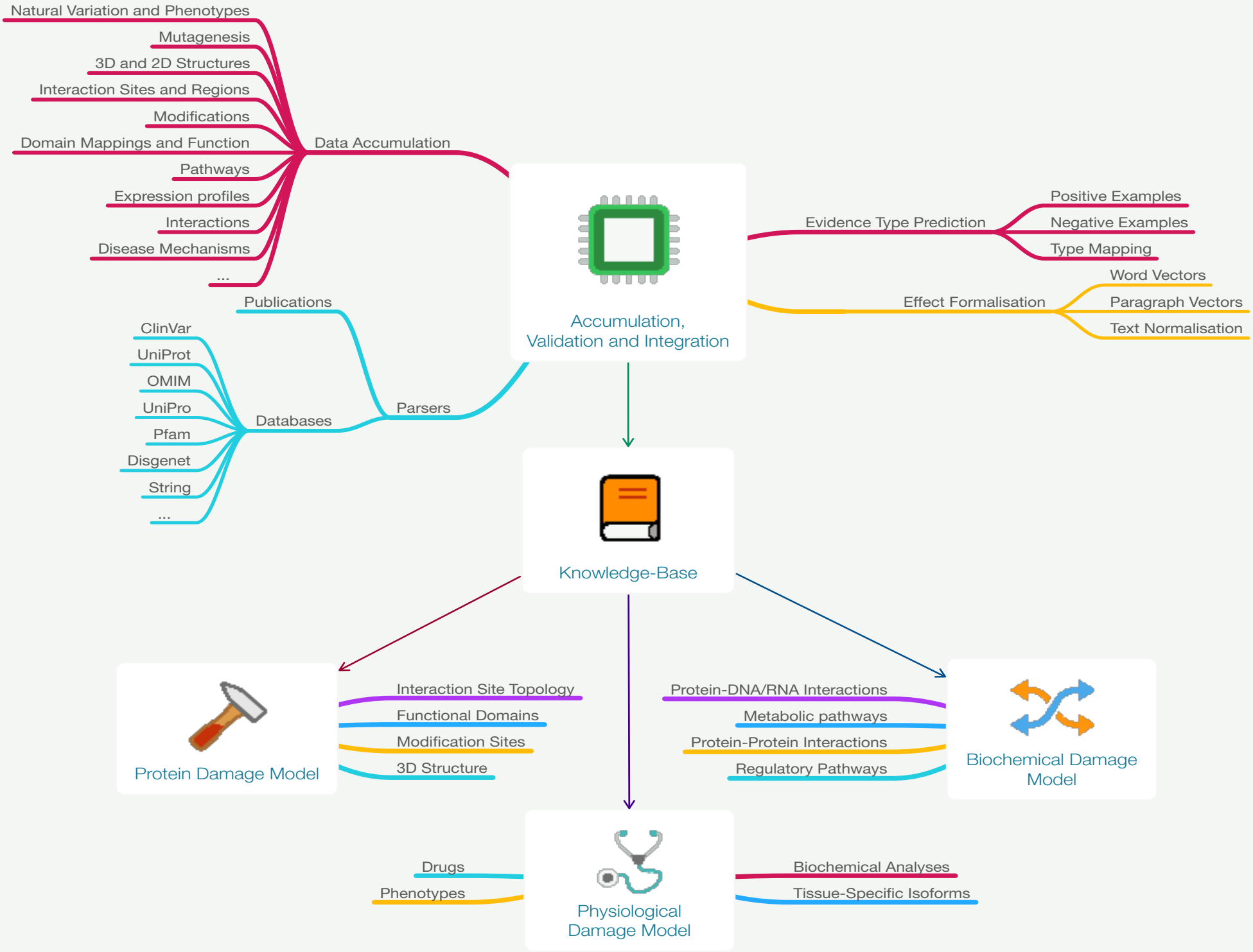
Научный руководитель:
Илья Корвиго
МФТИ
Лаборатория функционального анализа генома

Данные о связи генотипа и фенотипа

- Существует множество баз данных со сведениями о корреляциях генотип-фенотип.
- Подобная взаимосвязь может устанавливаться различными способами:
  - Статистические методы
  - Эксперименты *in silico*
  - Молекулярно-биологический лабораторный анализ

Наша задача в проекте

Подготовить к автоматической классификации данные из научных статей и записей из баз данных, описывающие ассоциации между вариантами и фенотипом, возникающие по причине геномных вариаций (SNP). Провести автоматическую классификацию без учета семантики.

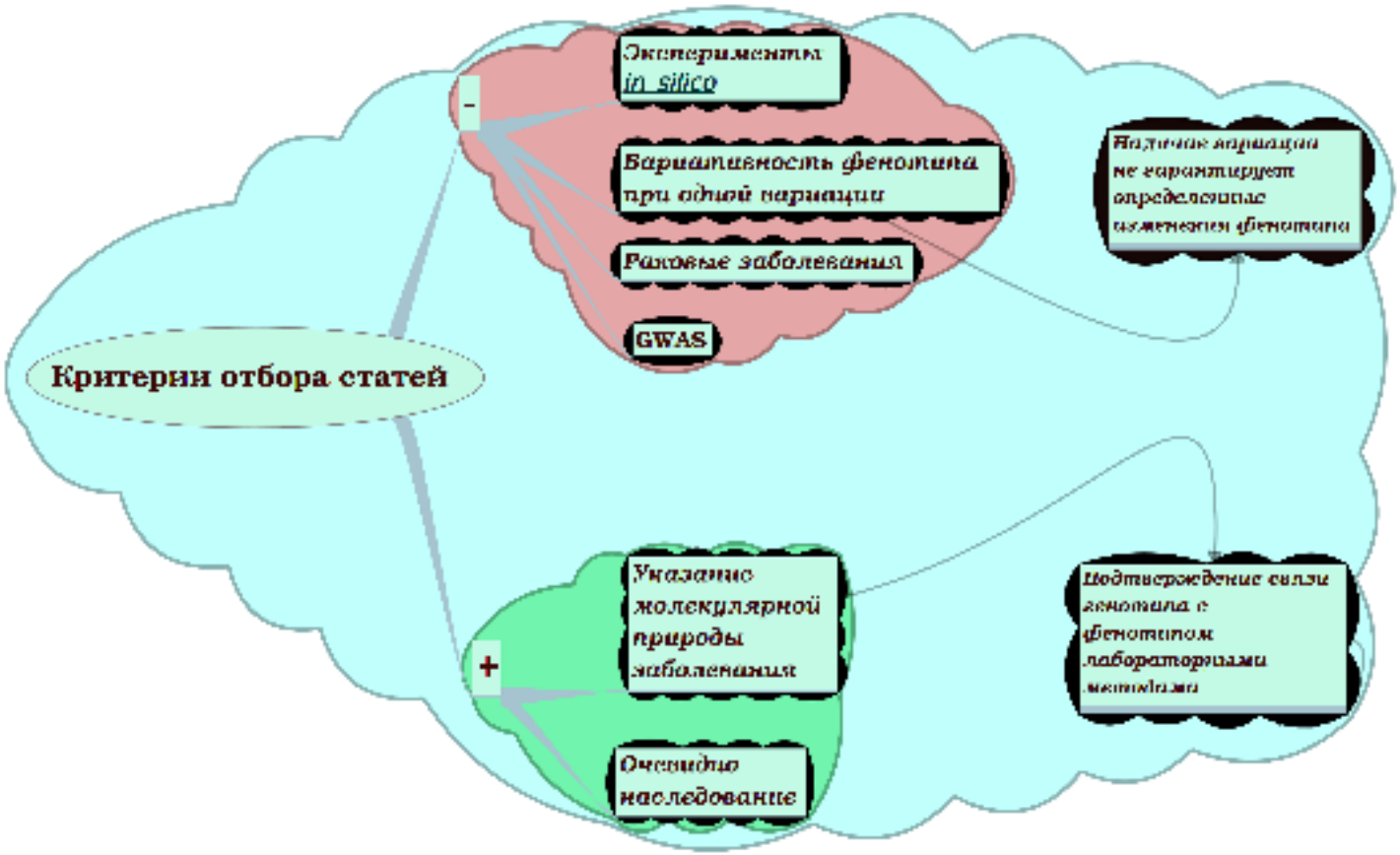

Natural Variation and Phenotypes
Mutagenesis
3D and 2D Structures
Interaction Sites and Regions
Modifications
Domain Mappings and Function
Pathways
Expression profiles
Interactions
Disease Mechanisms
...
Data Accumulation
Publications
ClinVar
UniProt
OMIM
UniPro
Pfam
Disgenet
String
...
Databases
Parsers
Accumulation, Validation and Integration
Evidence Type Prediction
Positive Examples
Negative Examples
Type Mapping
Effect Formalisation
Word Vectors
Paragraph Vectors
Text Normalisation
Knowledge-Base
Protein Damage Model
Interaction Site Topology
Functional Domains
Modification Sites
3D Structure
Biochemical Damage Model
Protein-DNA/RNA Interactions
Metabolic pathways
Protein-Protein Interactions
Regulatory Pathways
Physiological Damage Model
Drugs
Phenotypes
Biochemical Analyses
Tissue-Specific Isoforms

Подзадачи: Диана

- Парсер базы данных HUMSAVAR
- Отобрать из этой базы вариации, не относящиеся к полиморфизмам и раковым заболеваниям
- Найти пересечения вариаций из этой базы с вариациями из базы SwissProt, содержащей ссылки на статьи и «ECO»-коды для каждой вариации
- Распарсить 400 абстрактов, присвоив им «класс доверия»
- Обучить классификатор определять «класс доверия» статьи, провести 10-кратную кросс-валидацию

ИТОГ: кросс-валидация показала, что доля верных предсказаний составляет в среднем 66% и до 77% в отдельных сплитах.

Отбор статей

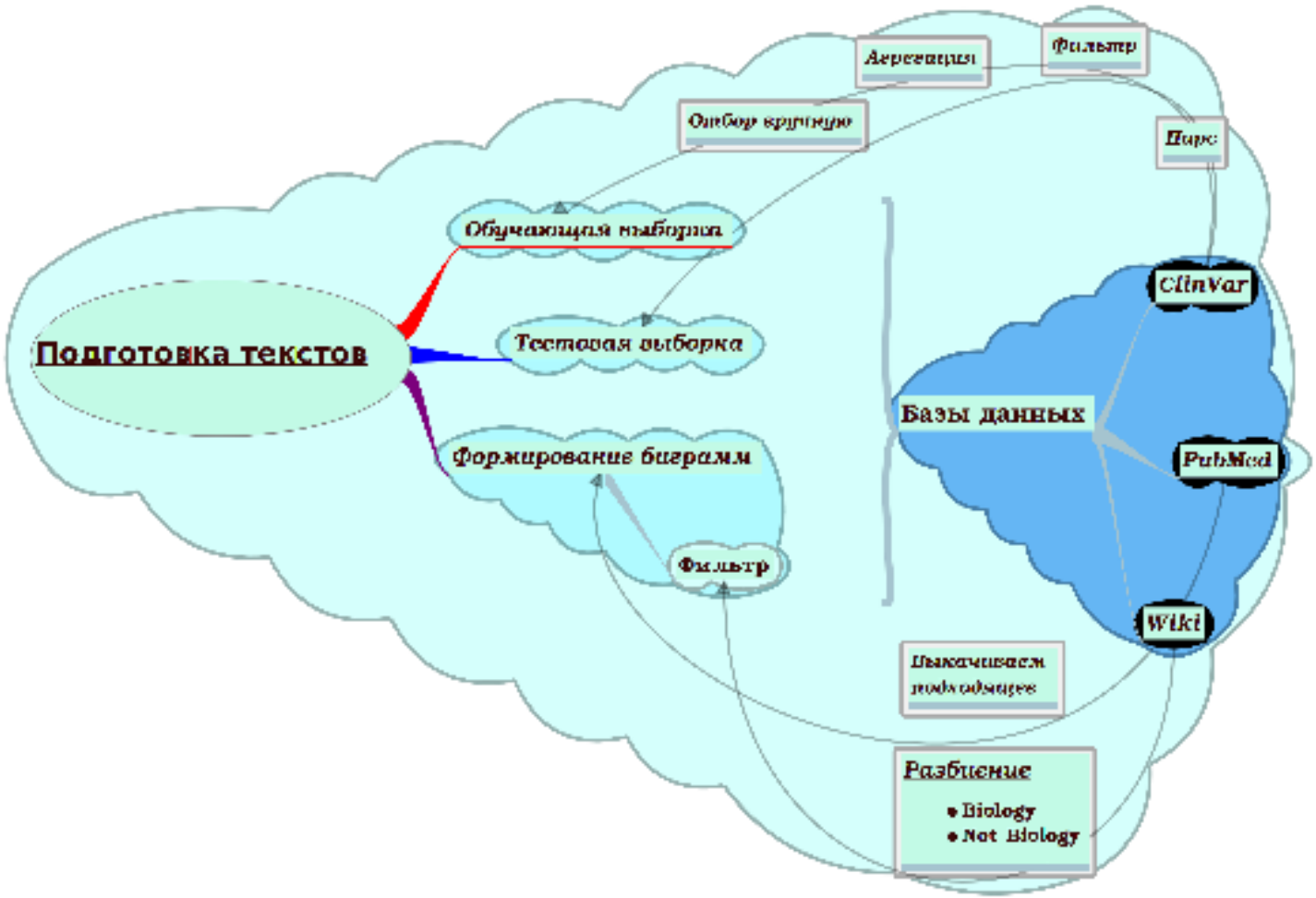


Подзадачи: Иван

- Парсер статей для создания биграмм
- Парсер базы данных ClinVar
- Агрегация статей по методу получения данных, указанных в этой базе
- Агрегация статей из Wikipedia по категориям на основе принадлежности к биомедицинским наукам при помощи pywikibot, парс этих статей при помощи pywikiextractor
- Агрегация абстрактов из PubMed на основе принадлежности к биомедицинским наукам и выделение из них биграмм

ИТОГ: собраны и подготовлены данные для создания моделей, описывающих влияние геномных вариаций на фенотип.

Агрегация статей из различных баз данных




THAT'S
ALL!