# Институт молекулярной генетики РАН

## Отдел молекулярных основ генетики человека

## Список тем студенческих работ

Предложенные ниже темы объединены общей научной задачей — выявления факторов риска возникновения ишемического инсульта головного мозга. Данный медицинский синдром проявляется в нарушении кровоснабжения отделов головного мозга в результате закупорки сосудов. Недостаточное кровоснабжение провоцирует нарушение функций клеток с последующим их отмиранием.

Ишемический инсульт (ИМ) является второй причиной смертности в Мире. Люди, пережившие инсульт становятся инвалидами, что приводит к социальным и экономическим последствиям. Уточнение существующих и выявление новых факторов риска позволит улучшить профилактику ишемического инсульта. Весомым результатом исследований является возможность сформировать рекомендации по профилактике рассматриваемого синдрома.

Факторы риска разделяют на изменяемые (воздействие среды обитания, образ жизни) и неизменяемые (пол, возраст, строение ДНК). Представленные ниже темы направлены на выявление факторов риска, обусловленных особенностями строения ДНК, а также клинико-биохимическими показателями.

Представленные ниже темы могут быть предложены студентам, специализирующимся в области компьютерных наук. Содержание тем, а также указанных в них сведений, является предварительным. Они могут быть скорректированы в ходе обсуждения с участниками исследований и дальнейшей проработки.

## Тема 1. Построение прогнозной модели с целью предсказания возникновения ишемического инсульта.

Дано: результаты генотипирования порядка 1000 больных ИМ и 1000 здоровых людей.

Требуется: выяснить возможность построения прогнозной модели с целью предсказания возникновения ИМ.

Данные генотипирования могут быть формализованы в виде объектно-признаковой матрицы, заполненной целочисленными значениями {0, 1, 2}. Они кодирую один из трех генотипов, возможных для данного признака. Размер матрицы 2000 x 80000.

Доступны также такие признаки как пол и возраст людей.

Целевая переменная: больной/здоров.

Ожидается что генотипические признаки будут ранжированы по вкладу в целевую переменную. Поскольку известно положение признака на ДНК, то можно строить рассуждения относительно влияния признака на организм человека.

Примеры из литературы. Используя методы машинного обучения, Kang MJ et al. (2019) построили прогнозные модели для предсказания размера ореха у каштанов (большой/маленький). Число генотипических признаков - 21. В обучающей выборке 42, в тестовой - 46 образцов.

## Тема 2. Импутация генотипов

Дано: 2-3 набора данных с разным числом генотипических признаков, например, 300K и 600K. Количество строчек в таблице 100 и 50, соответственно. Множества признаков плохо пересекаются между собой. Дан также референсный набор данных с числом признаков до 1M и числом строк до 1000.

Требуется: Повысить перекрываемость признаков при объединении изучаемых наборов данных. Для этого предлагается восстановить пропущенные признаки в изучаемых наборах данных, используя референсный массив генотипов.

## Тема 3. Выявление генетических факторов риска возникновения ишемического инсульта

Традиционный подход к выявлению участков генома, ассоциированных с заболеванием, — проведение полногеномного поиска ассоциаций (GWAS). В анализе обычно рассматривают три типа признаков (геномных маркеров):

1. Однонуклеотидный полиморфизм (SNP).

2. Гаплотип (комбинация SNP, наследуемых вместе).

3. Кластеры из гаплотипов.

Комбинации SNP можно определять с помощью методов машинного обучения (С. Sinoquet, 2018).

Требуется: Выяснить возможность применения методов машинного обучения для определения групп из SNP, а также провести GWAS с использованием данного геномного маркера.

## Тема 4. Прогнозировании исхода ишемического инсульта на основании клинико-биохимических показателей и полиморфизма генома больных инсультом.

Задача 1. Провести описательный анализ клинико-биохимических показателей больных с ишемическим инсультом с целью выявления типа переменных, полноты представленности данных, вариации значений, а также показателей, которые коррелируют между собой. Предложить способ нормирования значений. Анализ поможет определить множество переменных для решения задач, представленных далее.

Задача 2. Применяя методы машинного обучения, выполнить кластерный анализ клинико-биохимических показателей больных с ишемическим инсультом. Его цель — выяснить есть ли среди больных людей группы, чьи характеристики больше похожи между собой, чем между группами. Предложить различные меры сходства людей в группе, а также обоснование для выбора числа групп. Рассмотреть как иерархические, так и неиерархические модели кластеризации, как с четкими границами кластеров (один объект — один кластер), так и размытыми (один объект — много кластеров). Выполнить двумерное шкалирование.

Ожидаемые результаты кластерного анализа:

1. Понимание внутренней структуры данных

2. Возможность восстановления пропущенных значений клинико-биохимических показателей

3. Выявление типичного и нетипичного больного ишемическим инсультом. Нетипичный случай является выбросом и возможно этого больного нужно исключить из анализа.

Задача 3. Используя различные алгоритмы и методы машинного обучения, построить прогнозные модели, с целью предсказания развития, а также исхода ишемического инсульта.

В зависимости от задачи в качестве источника данных использовать базу данных больных с ишемическим инсультом, доступную в ОМОГЧ ИМГ РАН, а также данные генотипирования больных с ишемическим инсультом и здоровых людей.

## Данные

Предлагается использовать данные генотипов больных с ишемическим инсультом и здоровых людей, а также базу данных с фенотипами (пол, возраст), доступную в ОМОГЧ ИМГ РАН.

Прорабатывается также возможность использования данных из открытых источников, например, базы данных генотипов и фенотипов dbGaP (<https://www.ncbi.nlm.nih.gov/gap>).

## Литература

Christine Sinoquet, A method combining a random forest-based technique with the modeling of linkage disequilibrium through latent variables, to run multilocus genome-wide association studies. BMC Bioinformatics, 19:106, 2018. <https://rdcu.be/bxq5z>

Kang MJ, Shin AY, et al. Identification of transcriptome-wide, nut weight-associated SNPs in Castanea crenata. Sci Rep. 2019 Sep 11;9(1):13161. <https://www.ncbi.nlm.nih.gov/pubmed/31511588>

Vogel and Motulsky's, Human Genetics: Problems and Approaches. [Скачать pdf.](https://yadi.sk/i/N9Ye9OfRiYM9tg)

## Сокращения

GWAS — Genome Wide Association Studies (полногеномный поиск ассоциаций).

QTL — Quantitative Trait Loci (участки количественных признаков)

SNP — Single Nucleotide Polymorphism (однонуклеотидный полиморфизм)

ОМОГЧ ИМГ РАН — Отдел молекулярных основ генетики человека Института молекулярной генетики Российской академии наук

### Контакты:

Геннадий Хворых, гл.специалист, моб. 8-910-4389309, khvorykh@img.ras.ru